Tiago Miguel de Góis Raposo

Licenciado em Ciências da Engenharia Electrotécnica e de Computadores

C-RAN CoMP Methods for MPR Receivers

Dissertação para obtenção do Grau de Mestre em

Engenharia Electrotécnica e de Computadores

Orientador: Luis Filipe Lourenço Bernardo, Professor Associado com Agregação,
Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa

Co-orientador: Rui Miguel Henriques Dias Morgado Dinis, Professor Associado com Agregação, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa

Júri

Presidente: Prof. Doutor João Francisco Alves Martins, FCT-UNL
Arguente: Prof. Doutor Paulo Miguel de Araújo Borges Montezuma de Carvalho, FCT/UNL
Vogal: Prof. Doutor Luis Filipe Lourenço Bernardo, FCT-UNL

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Dezembro, 2018

C-RAN CoMP Methods for MPR Receivers

Este documento foi gerado utilizando o processador (pdf)LaTeX, com base no template "novathesis" [1] desenvolvido no Dep. Informática da FCT-NOVA [2].
[1] https://github.com/joaomlourenco/novathesis [2] http://www.di.fct.unl.pt

Acknowledgements

I, Tiago Miguel de Góis Raposo, hereby address my gratitude to my *alma mater*: the Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa (FCT-UNL), where I have spent six of the most rewarding and fulfilling years of my academic experience. The hardships faced throughout the Master's degree, and all the Professors and workmates that I have encountered along this journey made me grow not only as a future Engineer, but most importantly, as a human being thirsty for knowledge.

Also, I would like to express my gratitute to Fundação para a Ciência e a Tecnologia and Instituto de Telecomunicações, for the funding from the VELOCE-MTC - UID/EEA/50008/2013 project which conceeded me the research grant for this thesis.

I would also like to mention all the help and guidance that Professor Luis Bernardo provided along this year, and even before during all the courses that he lectured and I attended, the Professor was already showing signs of not only immense scientific know-how and experience, but also sublime teaching skills and with great ease, was able to captivate young students like me to take interest in the telecommunications area. Besides great technical mastery of the subjects of my dissertation, his patience and assistance with all my doubts and struggles throughout the development of this thesis, are more than worthy of this tribute that serves as the most sincere token of my appreciation for all the work that he has done, and surely will continue to do, to serve the University at the highest level.

A special note of gratitude to all the friends that I have met during these six years, with a honourable mention to my fellow colleague Rahim Karim Shamsudin, that was also under the guidance of Professor Luis Bernardo, and with whom I have shared my doubts and questionings along the theoretical research and practical development of my work, since sometimes we could only rely on each other to debate our own difficulties and reach new solutions. Also to all the student staff on STB, I leave a warm thank you for the joyful, memorable and motivating environment provided throughout this year in research lab number 1.9 in the Department of Electrical Engineering on FCT.

And last but not least, to my Family, Parents, Sister and Friends, that since a young age recognised in me the creativity and potential that allowed me to take higher flights, and provided the confidence and emotional foundation through all the tribulations in my life, the greatest thank you of them all.

Abstract

The growth in mobile network traffic due to the increase in MTC (Machine Type Communication) applications, brings along a series of new challenges in traffic routing and management. The goals are to have effective resolution times (less delay), low energy consuption (given that wide sensor networks which are included in the MTC category, are built to last years with respect to their battery consuption) and extremely reliable communication (low Packet Error Rates), following the fifth generation (5G) mobile network demands.

In order to deal with this type of dense traffic, several uplink strategies can be devised, where diversity variables like space (several Base Stations deployed), time (number of retransmissions of a given packet per user) and power spreading (power value diversity at the receiver, introducing the concept of SIC and Power-NOMA) have to be handled carefully to fulfill the requirements demanded in Ultra-Reliable Low-Latency Communication (URLLC).

This thesis, besides being restricted in terms of transmission power and processing of a User Equipment (UE), works on top of an Iterative Block Decision Feedback Equalization Reciever that allows Multi Packet Reception to deal with the diversity types mentioned earlier. The results of this thesis explore the possibility of fragmenting the processing capabilities in an integrated cloud network (C-RAN) environment through an SINR estimation at the receiver to better understand how and where we can break and distribute our processing needs in order to handle near Base Station users and cell-edge users, the latters being the hardest to deal with in dense networks like the ones deployed in a MTC environment.

Keywords: 5G, C-RAN, Multi-Packet Detection, Iterative Block Decision Feedback Equalization, Coordinated Multipoint, Power-NOMA, Ultra-Reliable Low-Latency Communication, Machine Type Communication.

Resumo

O crescimento da quantidade de tráfego nas redes móveis devido à utilização cada vez maior de aplicações para comunicação do tipo máquina (MTC) acarreta uma série de novos desafios no encaminhamento e gestão de tráfego ao longo da rede se quisermos ter tempos de resolução eficazes, pouca utilização de potência (tendo em conta que redes de sensores são dimensionadas para durarem anos em termos de utilização de bateria) e comunicação extremamente fiável (Packet Error Rate baixo) segundo as exigências das redes móveis de quinta geração (5G).

Para lidarmos com este tipo de tráfego, várias estratégias podem ser postas em cima da mesa em termos de diversidade de uplink, onde graus de liberdade como espaço (vários receptores na rede), tempo (número de cópias por utilizador numa transmissão de um dado pacote) e espalhamento de potência (diversidade de valores de potência na chegada ao receptor, introduzindo o conceito de Power-NOMA) têm de ser minunciosamente coordenados para atingirmos resultados dentro dos padrões de Ultra-Reliable Low-Latency Communication (URLLC).

Esta dissertação, para além de ter em conta as restrições em termos de potência e capacidade de processamento do hardware do utilizador (User Equipment), tem também em conta um modelo de receptor iterativo IB-DFE que permite recepção multi-pacote para lidar com os tipos de densidade citados acima. Os resultados desta dissertação exploram a possibilidade de fragmentar o processamento numa rede de acesso rádio baseada em computação em núvem (C-RAN) integrada para diminuir tempos de resolução na rede, através de um estimador de SINR na recepção que possa indicar quando e onde tratar utilizadores mais perto ou na fronteira entre células, sendo estes últimos os casos mais complicados de lidar numa rede móvel densa como as redes MTC.

Palavras-chave: 5G, C-RAN, Multi-Packet Detection, Iterative Block Decision Feedback Equalization, Coordinated Multipoint, Power-NOMA, Ultra-Reliable Low-Latency Communication, Machine Type Communication.

Contents

List of Figures

List of Tables

Acronyms

AP Access Point.

BBU Baseband Unit.

BS Base Station.

CC Code Combining.

CoMP Coordinated Multi-Point.

C-RAN Cloud Radio-Access Network.

CR Cognitive Radio.

CSI Channel State Information.

DC Diversity Combining.

EC Equal Channel.

ECPR Equal Channel with Phase Rotation.

eNB eNodeB.

H-NDMA Hybrid-ARQ Network Diversity Multiple Access.

IaaS Infrastructure as a Service.

IB-DFE Iterative Block Decision Feedback Equalization.

ICI Inter-Cell Interference.

LTE Long-Term Evolution.

MAC Medium Access Control.

MIMO Multiple-Input Multiple-Output.

MPR Multi-Packet Reception.

MTC Machine Type Communication.

NDMA Network Diversity Multiple Access.

NOMA Non-Orthogonal Multiple Access.

NORA Non-Orthogonal Random-Access.

OFDMA Orthogonal Frequency Division Multiple Access.

OMA Orthogonal Multiple Access.

ORA Orthogonal Random-Access.

PBCH Physical Broadcast Channel.

PER Packet Error Rate.

PHY Physical Layer.

QoS Quality-of-Service.

RAN Radio-Access Network.

RRH Remote Radio Head.

SCMA Sparse Code Multiple Access.

SDN Software-Defined Networks.

SDN-C Software-Defined Networks Controller.

SIC Successive Interference Cancellation.

SINR Signal-to-Interference-plus-Noise Ratio.

SNR Signal-to-Noise Ratio.

SP Equal Channel with Shifted Packet.

TDMA Time Division Multiple Access.

UC Uncorrelated Channel.

UE User Equipment.

URLLC Ultra-Reliable Low-Latency Communication.

Chapter 1

Introduction

Due to the increasing traffic demand in modern wireless networks, new ways are being sought to improve the performance of the network and the radio-access schemes.

Machine Type Communication (MTC) is a technology that realizes a system of networks for collecting data from machines such as sensors and smart meters that are usually massively, densely deployed. Unlike current world-scale, human-centric networks, MTC are featured with the absence of direct human intervention and with a rapid increase in connections count. It is anticipated that MTC will expand to 2.1 billion connections by the year 2021. Cellular network operators are considering MTC services as one of the key new revenue-generating services[1]. While the financial motivation for the network operators to massively deploy MTC services is clear, there needs to be a technological and architectural framework that is not only cost effective but also highly flexible to support the unique features of MTC traffic[1].

Different radio access technologies or techniques like cognitive radio can in fact improve the efficiency of the wireless system, but might not scale with the requirements of MTC traffic, reduce costs in BS hardware, or improve power savings. There is obviously a need to reach a middle ground between capacity, power saving, cost and spectral efficiency in 5G, while improving overall performance and reducing latency.

Beyond 2020 mobile networks need to support a 1000-fold increase in traffic relative to 2010 levels, and a 10 to 100-fold increase in data rates even at high mobility and in crowded areas if current trends continue[2].

A cloud processing approach centralizing control-plane configuration in the radio-access, and making devices along the network "dumber-but-faster" with reduced to none control functionality, improves bandwidth and permits a larger capacity for wide-sensor networks. This new concept of centralized processing can be seen as a form of a new C-RAN.

It would be interesting also to scale radio resources and processing capacity in highly congested areas where coverage is essential in the architecture. For instance, coverage patterns in a cell can be sensitive to hour, location and population density (e.g. one residential BS can be idle while another BS in a business area could be experiencing congestion[3]) whereas in Long-Term Evolution (LTE) standards, a BS is designed for peak traffic leading to power waste and elevated deployment costs, as the network expands with an oblivious design with respect to the heterogeneous nature of traffic patterns.

Another relevant Radio-Access Network (RAN) technology is Coordinated Multi-Point (CoMP). In CoMP, various cells cooperate to mitigate and try to nullify Inter-Cell Interference (ICI) or to increase signal at cell-edge. The integration of CoMP in a centralized processing environment can be explored to enhance performance and improve the power efficiency of the network.

On top of these architectural and design modifications in the network through software development, change is also necessary at the physical layer using innovative techniques, such as Multi-Packet Reception (MPR). In MPR, transmissions in the same channel from different users are allowed and devices are prepared to de-multiplex the superimposed signals through Non-orthogonal Multiple Access (NOMA) schemes in downlink and uplink channels, improving the spectrum usage efficiency.

All of the above described techniques allow for a new improved network management and overall performance, and it is in the scope of this dissertation to formulate a possible solution that unifies all the technologies described with the goal of approximation to an essential set of 5G services destined to provide to the users fast and error-free communication called Ultra-Reliable Low-Latency Communication (URLLC)[4]. URLLC deals with the deployment of very low latency adapted to each individual application (e.g authonomous driving, remote surgery).

1.1 Objectives and contributions

This dissertation addresses the development of a C-RAN architecture to scalably implement the 5G URLLC service. The goal of the author was, to study the improvements that the Iterative Block Decision Feedback Equalization (IB-DFE) receivers bring to deal with low-latency (through MPR and Code Combining (CC)) and explore, investigate and measure the performance of uplink diversity schemes (Power-NOMA (P-NOMA), Successive Interference Cancellation (SIC) or Hybrid-CoMP) that, embedded in an integrated C-RAN environment allow for a more reliable, scalable and faster mobile network performance.

A calculated analysis of SINR values on IB-DFE receivers is provided, based on previous works, that evaluates an optimal value γ, which, for a configuration of 2 BSs and 2 power levels between high and low power users, allows the author to distribute the processing load througout the network. Besides the SINR threshold obtained, a study on three different CPU families (AMD Ryzen 7, Intel Core i7 and Intel Xeon) is provided to

the reader, for better understanding the scalability of the setup in the context where the code was run.

1.2 Dissertation Structure

The dissertation structure is organized as follows: Chapter 2 contains a literature review about related work. It refers to various previous works on the impact of the deployment of a C-RAN to attain URLLC requirements, and how this drastic change in the architecture will adapt to increasing traffic and power saving demands. The second half of the literature review deals with the numerous diversity schemes searched by the author such as spatial, time and power-domain diversity, providing some preliminary perspective on approaches like NOMA, SIC or Hybrid-CoMP, used throughout the dissertation, and also introducing other techniques like SCMA, that are not further deployed by the author, but, nevertheless, are useful to understand the numerous range of diversity schemes than can be applied to attain URLLC requirements.

Chapter 3 introduces the new 5G C-RAN architecture and approaches, the mathematical models of the IB-DFE receivers used in the tests throughout this dissertation, and presents the models of the diversity schemes used by the author on top of said IB-DFE receivers. The simulations results are then presented and compared for two types of topologies (one and two BSs). Retransmission count, SINR at the receivers, and computation time to solve the matrices of each scenario are the main metrics considered throughout the chapter to reach a γ SINR to ensure reception given a deployed power configuration and distribution of UEs in a radio network. Finally, Chapter 4 contains the work conclusions and contains possible future work that can be done by taking this dissertation as reference.

Chapter 2

Literature Review

2.1 C-RAN: a new design

Several requirements are mentioned as goals for a new random-access MTC network in [2]. Enhancements in system capacity, throughput, less MTC latency, improved inter-device connectivity, reduced operational costs and consistent Quality-of-Service (QoS) are some of the mentioned features that are desirable in future wireless communications. Some solutions are proposed for the architecture[2], which require higher spectrum bands and greater spectral efficiency. Using unlicensed spectrum or fragmented spectrum aggregation through subcarrier aggregation can be an option to achieve it.

C-RAN is a solution that decouples processing resources from BSs throughout the network reducing BS functionalities, and providing it with layer 1/layer 2 tasks only, leaving higher layer processing capabilities for a cloud that serves multiple BSs. In consequence, the connections between Remote Radio Heads (RRHs)[1] and data centers can be scaled according to each link latency requirements, adding to the system the flexibility of deploying big or small cloud data centers.

Also, through SDN and NFV[2], when a data center is unable to respond to a flash crowd, processing capabilities can be requested to other less congested data centers. In a concept called Infrastructure as a Service (IaaS)[3], any idle network can be transformed in a processing resource for other more populated areas in the system. Figure 2.1 shows the

[1] A remote radio head (RRH), also called a remote radio unit (RRU) in wireless networks, is a remote radio transceiver that connects to an operator radio control panel via electrical or wireless interface.

[2] Network Function Virtualization. The main idea behind NFV is the decoupling of network functions from the physical network equipment where they run on. This is achieved by removing their execution from specific hardware and, by means of virtualization, run on standalone hardware[5].

[3] Infrastructure as a service (IaaS) refers to online services that provide high-level APIs used to dereference various low-level details of underlying network infrastructure like physical computing resources, location, data partitioning, scaling, security, backup etc.

interactions between the C-RAN and how it modifies the control-user plane connections in a 5G network. In the next subchapter, some practical deployments of the above strategies are presented to illustrate the already on-going attempts for a new architecture for 5G.

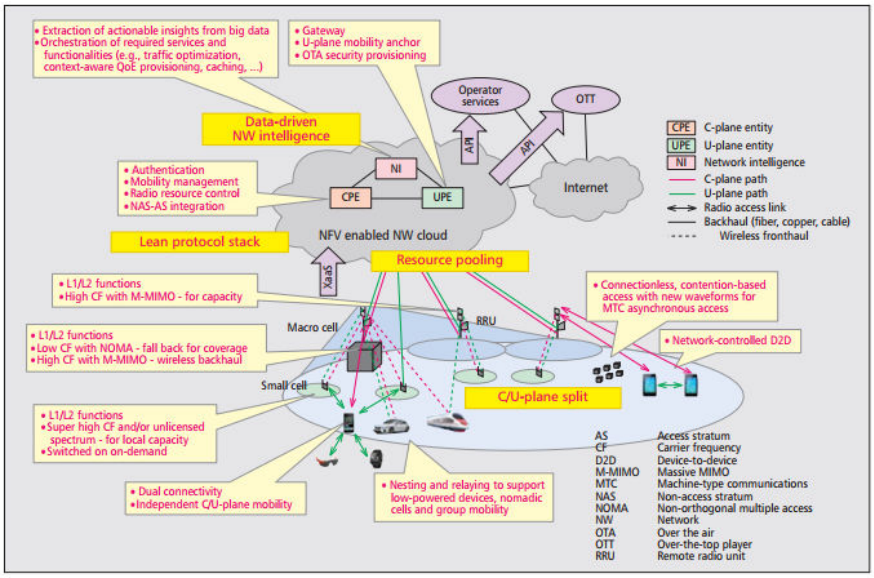


Figure 2.1: 5G mobile network vision and potential technology enablers[2].

2.1.1 Possible Solutions for 5G RAN

2.1.1.1 SDN in decoupling of control and user planes

Up until recent implementations of fourth generation mobile communication systems, each network element required individual data and control planes configuration. This methodology reduces the flexibility in the network due to the need to separately program each node using different management interfaces[6]. The potential of SDN is precisely the decoupling between the data plane (data forwarding) and control plane (management, configuration) router functions of each element. The control functions are aggregated in a software based controller which centrally, maintains an abstract perspective of the topology and provides a northbound configuration interface to the operators (figure 2.2) in order to, in a more efficient and faster way, impose and change data forwarding settings over each network router directly through the southbound interface (figure 2.2).

2.1.1.2 SDN handling MTC traffic

MTC defines a type of traffic with relatively low-rate and short-packet size. A scenario where MTC is used can be, for example, a sensor network over a railway, where various monitoring metrics are analyzed, but the sensors communicate between them sparsely over time and each transmission is minimized in terms of information. Rarely, bursts of information can occur during emergency situations.

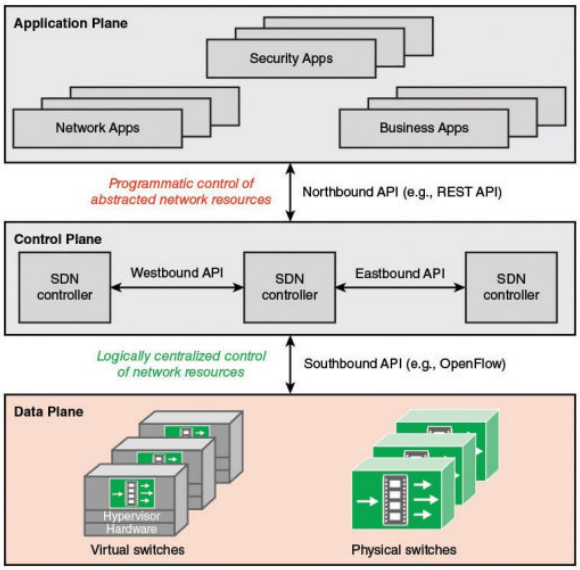


Figure 2.2: Between-layer SDN conceptual view

Correlation of MTC traffic (i.e. numerous sensors reacting with the same answers to the same event) and high densification of this type of traffic are two decisive factors that justify traffic aggregation of packets, in order to diminish the harmful effects of the traffic by requesting massive network resources and degraded performance that derives from the bursty nature of MTC.

The work in [1] displays an example of how network nodes can be centrally controlled by the Software-Defined Networks Controllers (SDN-Cs), thus, solving a traffic engineering [4] (TE) problem. From the individual network element informations and overall system requirements feedback, instructions can be complied by these nodes from the southbound interface in order to attain performance goals and optimize operations.

The decisions pertaining to traffic flow control and aggregation are made by the TE and traffic aggregation[5] (TA) modules interfaced to the SDN-C via the northbound interface. A two-phase traffic control (2PTC) mechanism is proposed, that takes advantage of the 5G architecture with pooled resources and centralized control. MTC packets are directed to virtual serving gateways (v-SGW) in the first phase of the proposal (M2GW phase). They are directed further on by traffic aggregation of correlated packets in a robust flow, to a sink in the second phase(GW2S phase).

Two problems are addressed in phase one: if the gateways are too close to the machines, they may not get enough packets to efficiently take advantage of traffic aggregation and, if gateways are too close to the sink and far away from hosts in the network it may cause sub-optimized M2GW communication. The work in [1] takes into consideration that the number of v-SGW's must be the smallest possible to reduce costs, as the correlated

[4] The atribuition of suited paths for the respective flows in the network.

[5] The aggregation of small time duration flows with the same sending and receiving pair and equal QoS requirements into a single long flow.

traffic must be aggregated to the fullest extent possible to increase packet compression.

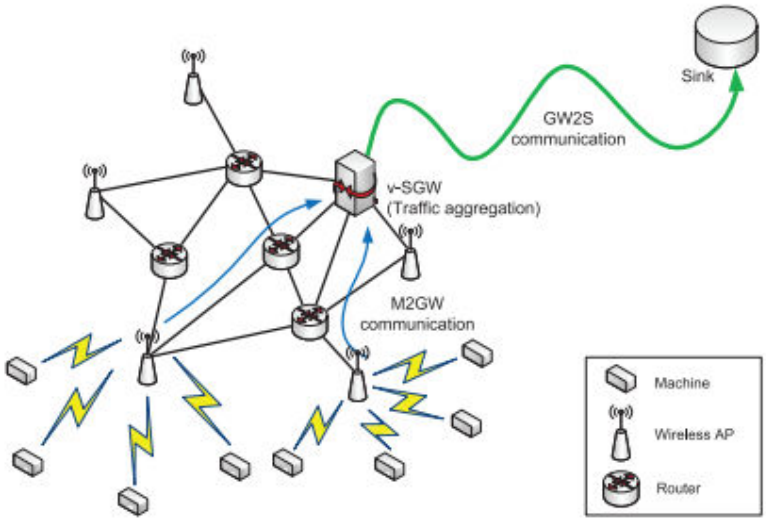


Figure 2.3: 2PTC for M2M communications[1].

Traffic aggregation

Some flows are so short in time duration that the optimization and decision process of where to send the packets throughout the network consumes more time than the travel itself. The solution would be to aggregate packets with the same source/destination pair while also being aware of similar QoS requirements between those packets.

In [1], a subset of network nodes are pre-configured to be v-SGW candidates. In each v-SGW there is a *f(k)* function that aggregates k packets to perform GW2S communication. Therefore, in this traffic control phase, slower rates are expected in the network due to aggregation in v-SGW's.

To calculate the correlation between packets, time is divided in time slots of equal length, and each slot is divided in p frames of equal length as well. For each machine m, the approach in [1] samples the average traffic rate $r_i^t(m)$ by time sub-slot i of each time slot t. The samples constitute a time series $r^t(m) = \left\{r_i^t(m) | i = 1, ..., p\right\}$ for every t. The term $\overline{r^t(m)}$ represents the mean of the sample set $r^t(m)$. In [1] the traffic rate correlation $\phi_{mm'}^t$ of two machines m and m' in time slot t is computed using the Pearson's correlation of the two sample sets $r^t(m)$ and $r^t(m')$ is measured as follows:

$$\phi_{mm'}^t = \frac{\sum_{i=1}^{p}(r_i^t(m) - \overline{r^t(m)})(r_i^t(m') - \overline{r^t(m')})}{\sum_{i=1}^{p}\sqrt{(r_i^t(m) - \overline{r^t(m)})^2}\sqrt{(r_i^t(m') - \overline{r^t(m')})^2}} \tag{2.1}$$

The correlation $\phi^t_{mm'}$ is represented by two values, -1 and 1. -1 means that high (low) scores on one variable are assigned to relatively high (resp. low) scores on the other. On the other hand, -1 means that high (low) scores on one variable are mapped to relatively low (resp. high) scores on the other. In [1] the correlation distance between m and m' is formulated as

$$d_{mm'} = 1 - \widehat{\phi_{mm'}} \tag{2.2}$$

where $\widehat{\phi_{mm'}}$ is the expected correlation value in the most recent q time slots.

Assigning machines to virtual gateways

Three goals are defined in order to optimize v-SGW's-to-hosts assignment in the network:

- Minimize the average assigning cost from the network resource usage point of view (Assigning Cost Minimization).
- Minimize the number of used v-SGW's in the network from the operational cost point of view (Gateway Selection Minimization).
- Minimize the correlation cost of same v-SGW's machines to maximize the payload aggregation potential (Correlation Distance Minimization).

The first goal problem size depends on the number of machines in the network such that, if it is too large, may cause delay in the convergence of the solution. Therefore, the clusters are formed according to the per-machine density, which consists in the ratio between machine traffic and total traffic in the network. In [1], clusters are processed from the largest to the smallest density regions.

2.1.1.3 Network slicing applied to power saving

A RANs traffic changes significantly throughout the day, and the digital unit of a BS is designed for peak traffic. Due to the non-constant traffic usage pattern in a network, and knowing that each BS needs housing facilities (e.g. cooling) that consume constant power, we might have considerable unnecessary power wasting by the deployed idle BSs during the day. One possible solution to adapt the network to the user pattern, could be the virtualization of processing resources for each BS, making it available and shareable between several cells. In [7] is shown that, by dynamically assigning virtual resources to each cell in the form of a virtual fronthaul link combined with virtual functions that perform baseband processing in the cloud, we are enabling each cell with a virtual base station (VBS). This VBS can be controlled by a SDN-C that decides to which cell belongs each VBS and how to form one depending on real-time statistics of traffic charge.

2.1.2 Issues with the clouded approach: Ultra Reliable Low Latency

2.1.2.1 Impact of C-RAN in URLLC

In order to attain URLLC requirements introduced in 1, three problems can be identified in order to reach more latency-sensitive solutions: overhead relief, which pertains to need to diminish the size of information like channel training, resource allocation or user scheduling; packet error probability decrease, since URLLC no longer withstands retransmissions, therefore, the PER of the first transmission must drop, and finally, the delay of this first transmission must also diminish.

The features of a C-RAN implementation clash with a reliable and fast communication philosophy in several aspects:

- In resource optimization, the computational complexity at the CPU pool rises with the number of devices, radio resources and eNodeBs (eNBs) which will consequently impact the delay.
- Unacceptable overhead signaling in the air due to access control policies to properly assign resources has a cost on latency requirements as well.
- Data routing and paging decisions take its toll on network time resources due to their non-dynamic configuration in the network, which means that the network assumes that the device can easily reach any part of the world ignoring a certain tendency in user patterns to only communicate with a restrained number of devices throughout the net (social profile of the device).

Hence, three latency types have been identified: radio-access latency, resource optimization latency, and routing/paging latency. Each requires special handle in order to mitigate them in the network.

2.1.2.2 Possible solutions for URLLC in C-RAN

Given the latency types presented in the former section, [8] proposes three possible solutions for every one of them.

Open-loop communication

In 4G and 3G, parameters such as channel estimation, data transmission schemes or power control are only decided after feedback from the receiver has been sent. In an open-loop paradigm, there is no optimization of transmission parameters due to the lack of receivers' feedback. The transmitter determines the transmission technique (modulation and code scheme), space-time code for Multiple-Input Multiple-Output (MIMO) and retransmission number in time, frequency and space for the first try. Medium access is open-loop and therefore, medium-sensing is applied (e.g. Cognitive Radio (CR)) to avoid interference.

Information-bridled resource optimization

To avoid the signaling overhead in the air due to constant exchange of information like Channel State Information (CSI), a transmitter only needs the CSI, interference levels and transmission schemes statistics in a long-term period. Such information can be globally provided for all devices in the network thus enabling each transmitter to optimize its transmission scheme. The C-RAN broadcasts the configuration statistics to all devices implicitly controlling their radio accesses.

Social data cache based paging/routing

In order to shorten the exchange of paging and routing steps every time a device wants to communicate, new sequences of events must be devised by the architecture and thereby, reduce latency.

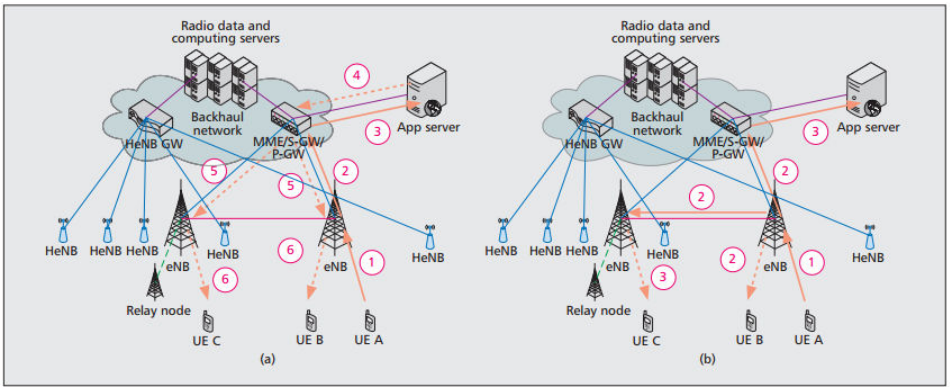


Figure 2.4: a) The amount of steps needed to exchange a message between two terminals; b) by memorizing social patterns of UEs in the C-RAN, latency in the network can be significantly reduced[8].

The approach in [8], which corresponds to the scheme in figure 2.4b), is composed by:

1. UE-A sends a message to eNB.

2. If the C-RAN already saves in a cache the information about the message destination, the eNB sends this message directly to UE-B, and, at the same time passes it to S-GW/R-GW.

3. eNB sends the message to UE-C and S-GW/P-GW passes the message to the app server.

The approach in figure 2.4a) ignores the social correlation between each device in terms of geographical location, identity, etc. Such social profile of each device can remain unmodified and therefore form a solid indicator about the routing dynamics in the network.

2.1.3 CoMP in a C-RAN environment

CoMP underlines a set of techniques for neighboring cell cooperation to mitigate interference and improve the signal for cell-edge users. In [9], northbound algorithms to the SDN-C are applied to coordinate users for downlink and uplink CoMP methods. Two essential issues are tackled: the uplink coordinated user selection and the downlink ICI cancellation.

Uplink coordinated user selection

1. Each UE transmits towards the RRH of the cell to which it is associated and single-user detection is performed in the Baseband Unit (BBU)[6] (eNB).

2. After Physical Layer (PHY) processing, error rates and received signal characteristics (modulation order, received power) are sent to the coordinator.

3. Coordination algorithm detects high ICI and enables multi-user detection. It sends to the eNBs scheduling constraints for users involved in joint detection and instructions to activate joint detection functions (i.e., multi-user channel estimation, MMSE(minimum mean square error) matrix computation and MMSE equalization).

4. UEs transmit according to new parameters.

5. Multi-cell joint MMSE detection is realized. Error rate is decreased with regards to previous transmission and effective throughput is improved.

Downlink inter-cell interference cancellation

1. DPB (dynamic point blanking). The process of identifying interferers in downlink direction of a given UE B. By muting the dominant interferer, the SINR can be improved.

2. Defining neighboring cell. When the difference between the average power received from the neighboring cell, denoted by P_{uc}, and that received from the serving cell, denoted by P_{us}, is less than a given predefined threshold dP, the neighboring cell is defined as a cooperating cell for the considered user. Thus, the set of cooperating cells C_u of user u is updated over time based on long-term UE power measurement. It is not expected to change over time if the location of the UE does not change.

3. Select a scheduler that has the knowledge of CSI info of all users being served within the cluster based in instantaneous data rate relative to its mean data rate.

[6] A baseband unit (BBU) is a unit that processes baseband in telecomm systems. The baseband unit is placed in the equipment room and connected with RRU via optical fiber. The BBU is responsible for communication through the physical interface.

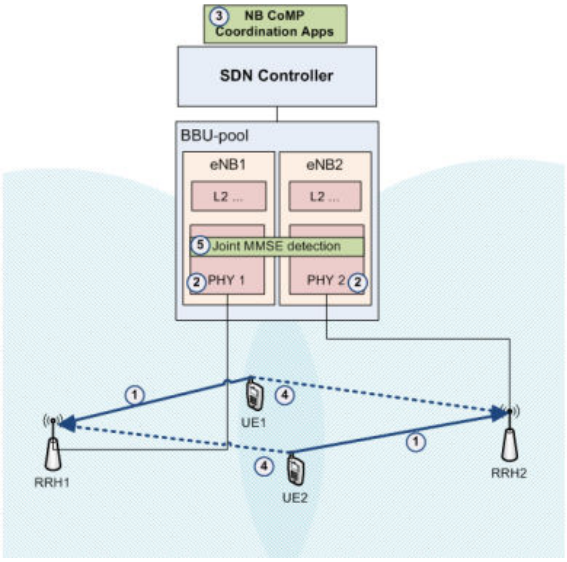


Figure 2.5: Uplink joint detection architecture and coordination process[9].

What is evaluated in [9] is the mean user throughput with or without multi-cell interference cancellation, throughput gain with increasing transmission power with dynamic cancelation of interference, and mean number of users per cell with increasing power.

The work in [9] evaluates latency times in both northbound and southbound interfaces. The flow of information monitoring contemplates the time under which metric updates are available to the northbound API's which then take action through their algorithms, thus renewing the output configuration parameters targeting the southbound interface - which speaks directly with the network nodes - with the same protocols.

The investigation in [9] concludes that the traffic times range from $0.4ms$ to $0.7ms$, which means that it is expected a delay of $2ms$ between the capturing of new measures and the updating of such values. It also confirms that multi-cell coordination can be verified with up to $300km/h$ of user velocity. This could support users with high mobility.

2.2 Embracing the collision

The collision model, which is based on the assumption that when only one user transmits the packet reaches the receiver error free, but when more than one user transmits data, the packets at the receiving end are dropped due to collision.

Of course, this approach lies in two extremes of the network transmission scenarios, neither all one-user exclusive transmissions arrive error free due to fading or noise at reception, neither multiple user transmissions can be dismissed with such ease due to strategies like code combining employed in multi-packet reception schemes [10].

Until recently, the theory of random-access was based on such an idealized model, and random-access protocols were viewed as collision resolution or collision avoidance techniques. In practice, the collision model is both optimistic and pessimistic: optimistic,

for it ignores channel effects such as fading and noise on reception, and pessimistic, because it does not accommodate the possibility that packets may be successfully decoded in the presence of simultaneous transmissions[10].

According to [11], theoretically, orthogonal transmission is suitable for downlink as it can achieve the maximum users' sum-rate. In uplink, orthogonal transmission is not optimal in terms of spectral efficiency, and cannot achieve the system upper bound for delay-sensitive applications. There is, therefore, a necessity to get rid of orthogonality of access in traditional multiple access schemes where each user is given a resource block (time, frequency or code), which is commonly designated as Orthogonal Multiple Access (OMA).

Such a shift in the access paradigm sheds light over a new area of research in wireless communications which embraces collision and non-orthogonal access, thus new acronyms like NOMA, MPR, and SIC are now part of the vocabulary.

2.2.1 Successive Interference Cancellation (SIC) and the need for power allocation in throughput and spectral efficiency

New ways are being sought in the literature and in practical deployments to handle collision more efficiently spectral and delay-wise, rather than discarding all the collided packets and demand retransmission. In uplink or downlink scenarios, the devices must be equipped with receivers applying multi-user detection techniques and access schemes that allow for superposition of signals from various users over the same resource.

The basic NOMA scheme applying SIC for UE receivers in the cellular downlink was introduced in [12]. For simplicity, it is assumed a two UE case, a single transmitter, and a single receiver antenna. The BS transmits a signal for SIC-i ($i = 1, 2$), x_i, where $\mathrm{E}\left[|x_i|^2\right] = 1$, with transmission power P_i. The sum of P_i is restricted to P at maximum. In the NOMA, x_1 and x_2 are superposition coded as:

$$x = \sqrt{P_1}x_1 + \sqrt{P_2}x_2 \tag{2.3}$$

The received signal at UE-i is represented as:

$$y_i = h_i x + w_i, \tag{2.4}$$

where h_i is the complex channel coefficient between UE-i and the BS. Term w_i denotes the receiver Gaussian noise including ICI. The power density of w_i is $N_{0,i}$. From transmission to reception: the signals are organized first in the downlink by the BS transmitter linearly adding them up in assigned power partitions to even the sum rate of all the users under a given power subset, and to maintain throughput fairness between individual users.At reception, SIC is employed to perform multi-user detection (MUD). Some channel condition scenarios are not ideal for SIC, such the ones provoked by the near-far effect. In

general, SIC is used at users with high SINR, and is carried out from highest to lowest values of received power. A similar scheme can be used for uplink to increase the uplink system capacity[13]. The throughput of UE-i, R_i, is represented in [12] as:

$$R_1 = \log_2\left(1 + \frac{P_1|h_1|^2}{N_{0,1}}\right), R_2 = \log_2\left(1 + \frac{P_2|h_2|^2}{P_1|h_2|^2 + N_{0,2}}\right) \tag{2.5}$$

From (2.5) we can see the purpose in adjusting power allocation configuration in order to modify modulation, coding scheme and throughput all over the topology during uplink transmission. The power ratio $\frac{P_1}{P_2}$ can be used by the BS to control the throughput of the receiving UE, thus allowing the BS to, more efficiently approach total fairness among users and flexibly utilize power diversity in the radio interface.

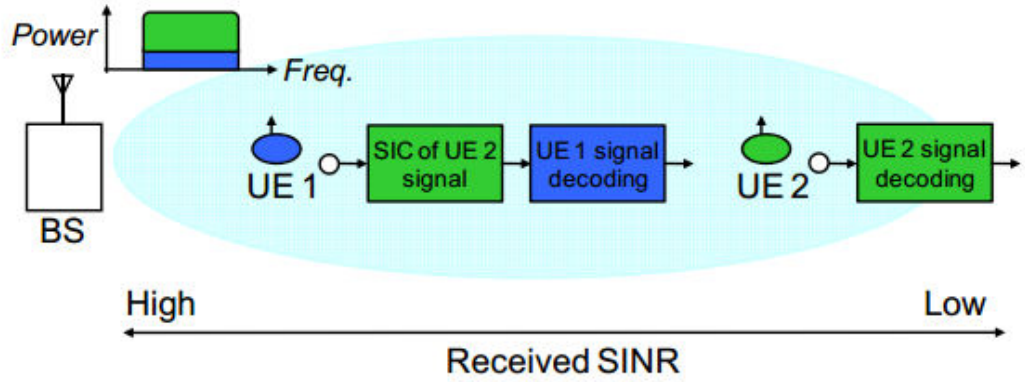


Figure 2.6: Basic NOMA scheme applying SIC for UE receivers in downlink[12].

To highlight the relevancy of fine power allocation schemes in QoS for downlink and uplink NOMA systems, the work in [14] devises a new scheme called D-NOMA (dynamic NOMA) which is built on top of two initial restrictions (both for downlink and uplink, totalizing 4 restrictions overall) that guarantee that the individual rates in the network are better than those in OMA. In this article a system with M single-antenna users and one single-antenna BS is considered. It is assumed that the users are ordered as $|h_1|^2 \leq |h_2|^2 \leq \cdots \leq |h_M|^2$, where h_i is the Rayleigh fading channel gain.

The different choices in v and ω, as [14] comes to show, may impact on the use of power diversity (NOMA), which in turn, will affect the way users are clustered throughout the topology. NOMA can be used for each predetermined subset of users in a cell, each subset being distinguished by some other diversity layer on top on NOMA (e.g different n subcarriers or n time slots for different n-NOMA groups). The data rates of user v and ω in downlink NOMA can be given by [15]:

$$R^N_{v,D} = \log_2\left(1 + \frac{\alpha_v|h_v|^2}{\alpha_\omega|h_v|^2 + 1/\rho}\right), \tag{2.6}$$

and

$$R^N_{\omega,D} = \log_2\left(1 + \alpha_\omega\rho|h_\omega|^2\right), \tag{2.7}$$

respectively, where ρ is the transmit Signal-to-Noise Ratio (SNR), α_v and α_ω are the power allocation factor for user v and ω, respectively, and $\alpha_v > \alpha_\omega$, $\alpha_v + \alpha_\omega = 1$. $R^N_{\omega \to v,D}$ denotes the rate for user ω to decode user v's message, i.e.,

$$R^N_{\omega \to v,D} = \log_2\left(1 + \frac{\alpha_v \rho |h_\omega|^2}{\alpha_v \rho |h_\omega|^2 + 1}\right). \tag{2.8}$$

Since channel power is always superior for user ω in comparison to user v, user ω can always detect user v's message before decoding its own. As a consequence the condition $R^N_{\omega \to v,D} > R^N_{v,D}$ always holds. By contrast, when OMA is used for user i applying time diversity Time Division Multiple Access (TDMA), the rate R_i is given by:

$$R^T_i = \frac{1}{2}\log_2\left(1 + \rho |h_i|^2\right), i \in \left\{v, \omega\right\}. \tag{2.9}$$

The work in [16] describes NOMA as a special case of a CR system. In [14], the user ω with the best channel conditions is labeled as a primary user, and the rate at said primary user ω is $R^D_{\omega,N} \geq R^T_\omega$, resulting in the following condition:

$$\begin{aligned} \log_2\left(1 + \alpha_\omega \rho |h_\omega|^2\right) &\geq \frac{1}{2}\log_2\left(1 + \rho |h_\omega|^2\right) \\ \Rightarrow \alpha_\omega &\geq \frac{1}{\sqrt{1 + \rho |h_\omega|^2} + 1} \end{aligned} \tag{2.10}$$

In addition, [14] assumes that user v with poor channel condition can also be regarded as a primary user, and assume that the target rate at user v is R^T_v , which means $R^D_{v,N} \geq R^T_v$, then

$$\begin{aligned} \log_2\left(1 + \frac{\alpha_v |h_v|^2}{\alpha_\omega |h_v|^2 + 1/\rho}\right) &\geq \frac{1}{2}\log_2\left(1 + \rho |h_v|^2\right) \\ \Rightarrow \alpha_\omega &\leq \frac{1}{\sqrt{1 + \rho |h_v|^2} + 1} \end{aligned} \tag{2.11}$$

The upper limit of α_ω, satisfies $1/\sqrt{1 + \rho |h_v|^2} + 1 < \frac{1}{2}$, which contributes for ensuring that user v increases its transmission power [14],[15],[17].

Power constraints are essential in dimensioning a network under some practical scenarios. For a cell with various users adopting a shared bandwidth, the transmission power constraint might prove crucial to manage ICI.

For uplink NOMA, the work in [14], surmises that the order of decoding occurs always from best to worst channel conditions. From the other way around, a substantial amount of transmission power needs to be consumed by user v to balance the channel attenuation.

The rates of user ω and user v are given by:

$$R^N_{\omega,U} = \log_2\left(1 + \frac{\alpha_\omega |h_\omega|^2}{\alpha_v |h_v|^2 + 1/\rho}\right), \tag{2.12}$$

and

$$R^N_{v,U} = \log_2\left(1 + \alpha_v \rho |h_v|^2\right), \tag{2.13}$$

respectively.

Similarly to downlink NOMA, [14] firstly considers the constraint $R^N_{\omega,U} \geq R^T_{\omega}$, which yields the following:

$$\begin{aligned} \log_2\left(1+\frac{\alpha_\omega|h_\omega|^2}{\alpha_v|h_v|^2+1/\rho}\right) &\geq \frac{1}{2}\log_2\left(1+\rho|h_\omega|^2\right) \\ \Rightarrow \alpha_\omega &\geq \frac{\left(1+\rho|h_v|^2\right)}{1+\rho|h_v|^2+\sqrt{1+\rho|h_\omega|^2}}. \end{aligned} \tag{2.14}$$

Secondly, [14] considers that $R^N_{v,U} \geq R^T_{v}$, which leads to the following:

$$\begin{aligned} \log_2\left(1+\alpha_v\rho|h_v|^2\right) &\geq \frac{1}{2}\log_2\left(1+\rho|h_v|^2\right) \\ \Rightarrow \alpha_\omega &\leq \frac{\sqrt{1+\rho|h_v|^2}}{\sqrt{1+\rho|h_v|^2}+1}. \end{aligned} \tag{2.15}$$

Using these restrictions (for downlink in (2.10) and (2.11) and uplink in (2.14) and (2.15)) an expression can be built for the power coefficients, as these dynamically change with the channel gains. Based on these coefficients, the expressions for individual rates can be obtained and used in performance analysis. Rate probability and average rate are used as criteria to analyze the performance of the proposed scheme with dynamic power allocation.

The paper in [14] compares the proposed dynamic power allocation strategy (D-NOMA) with the following alternatives:

- F(Fixed)-NOMA, where the power coefficients are fixed and not a result of a function of the channel gains. The disadvantage of this proposal is that the predefined quality for the users is simply not reached in practical scenarios. In comparison with OMA the same weak channel user has a much lower rate in F-NOMA, especially in a high SNR environment.

- CR-NOMA, where the weak channel user is prioritized and its QoS demands are firstly met. The drawback here happens when this weak user demands higher and higher rates which, in turn, will require high transmission power to enlarge its SNR, thus draining all the power that otherwise would be employed in a better channel user, leaving the later underserved.

The results in [14] demonstrate that the diversity gain (increase in the SNR due to some diversity scheme) in D-NOMA is similar to F-NOMA but D-NOMA achieves a better balance in user service. This means that D-NOMA can avoid the situation where the weak channel user is served with fewer rates, comparing with F-NOMA. The diversity gain of D-NOMA also outperforms the gain in CR-NOMA.

In [18], a decentralized transmission power control scheme is applied to a two-user scenario where every user chooses its transmitted power level in a random fashion, in

agreement with a given power distribution. The scheme suits notably for a cognitive radio system with a number of second-tier users transmitting whenever the oportunity arises over spectrum holes. In this kind of system, the opportunity that the spectrum holes bring might be outweighed by the overhead necessary to establish centralized control, as a spectrum hole may last only a short burst of time.

2.2.2 Multi-user detection in power domain impacting random-access

MTC communication implies massive access, which implies collision in practice, causing congestion, which ends in delay or impossibility of accessing the network. In [19], a NORA scheme is proposed to deal with the increasing number of UEs in a MTC scenario.

As expected, more user density means more collision which leads to congestion. The amount of overhead per UE keeps increasing until the maximum preamble number allowed per user is reached. After that the UEs give up on the random-access procedure and acknowledge its failure. Even if preamble maximum is not reached, the access delay will be unbearable afterwards if the UE succeeds to complete the random-access process. As a result, the blocked access will lead to unutilized resources in the network since UEs cannot overcome the first access step.

In contrast with the orthogonal scheme (ORA), NORA eases the simultaneous transmission of Msg3 in figure 2.7 of collided UEs as opposed to leading to retransmission of access overhead (preambles), which bypasses future increasing collision and prevents demands on PUSCH resources in the process. To conduct NORA multiplexing of users and SIC on the BS side, UE location and channel condition information is also used.

Usual power control strategies try to preserve a constant received power at the BS coming from various UEs.

In NORA, the BS performs user separation based on SIC, which requires diverse arrived power from the UEs. The transmit power of the i-*th* UE in a NORA group is expressed by:

$$P_{U,i} = \min\left\{P_{Umax}, P_{O_U} - (i-1)\delta + 10\log_{10}\left(M_{U,i}\right) + \alpha P_L i\right\}, \tag{2.16}$$

where δ is the Power Back-off Offset broadcast on Physical Broadcast Channel (PBCH), P_{Umax} is the maximum transmit power and P_{O_U} represents the received power per resource block when assuming a path loss of $0dB$. M_U denotes the number of available resource blocks in UL grant while P_L denotes the downlink path loss estimate, α represents the reduced rate of transmit power increase due to fractional power control.

The power back-off order of UEs in a NORA group is decided according to the TA[7] (timing advance calculated based on the delay spread[8]of the preamble) value. The UE

[7]Timing advance value corresponds to the length of time a signal takes to reach the base station from a device.

[8]Difference between the time of arrival of the earliest significant multipath component (typically the line-of-sight component) and the time of arrival of the latest multipath components.

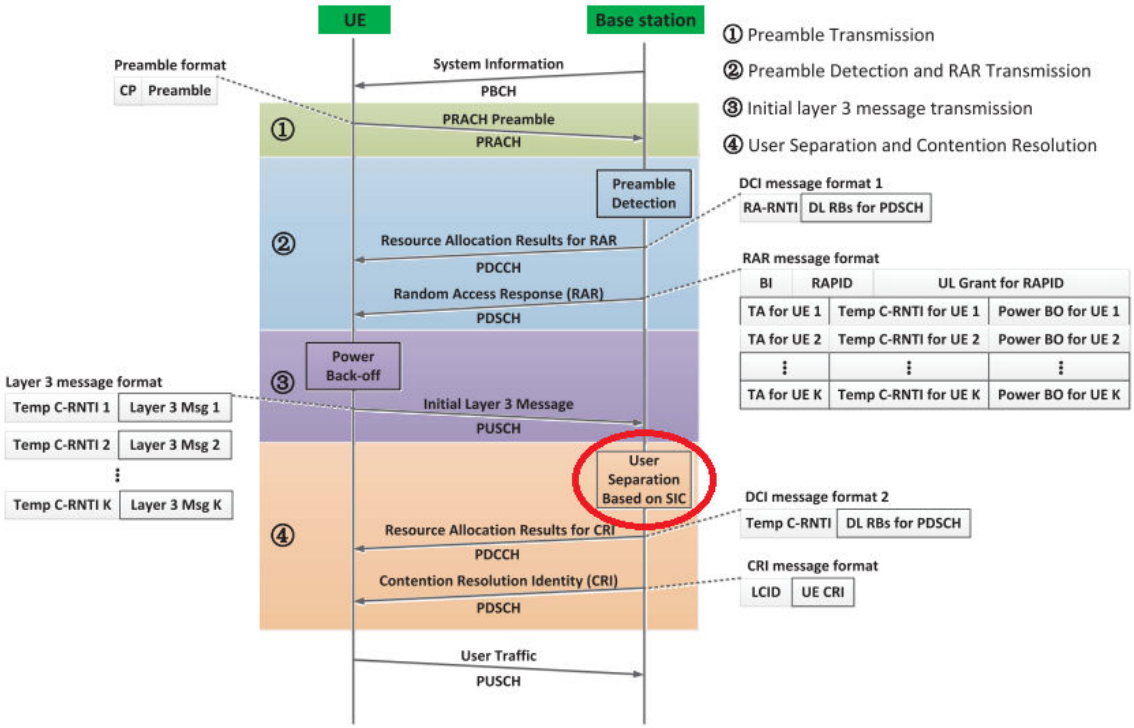


Figure 2.7: Non-orthogonal Random-Access Process[19].

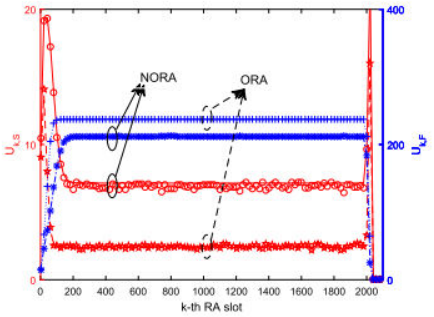
with a larger TA will be assigned a larger order i, which indicates that the corresponding received power $P_{U,i}$ is smaller.

The decoding order of UEs in a NORA group is consistent with the power back-off order, i.e. the UE with the strongest received power will be decoded first.

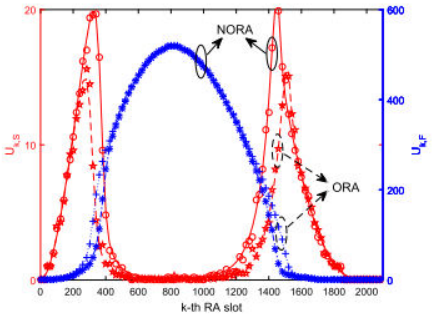
Two traffic models are simulated in [19], where the figure 2.8 shows the number of succeeded UEs $\left(U_{k,MS} = \sum_{l=1}^{L} U_{k,MS}\right)$ and failed UEs (U_k - $U_{k,MS}$) in the k-*th* RA slot of the NORA and ORA schemes under Traffic Model 1 and 2 (figures 2.8 a) and 2.8 b), respectively) $U_k = 50000$ is taken to model the overloaded scenario.

Figure 2.8(a) shows the comparison between ORA and NORA approaches. It is visible that ORA and NORA reach a peak at the start of the random-access process. Afterwards, ORA performance begins to fall with respect to the number of succeeded UEs. When performance is steady, the non-orthogonal approach is threefold better than ORA, probably due to the non-orthogonal stacking of preambles and message procedures during random access phase. Moreover, ORA experienced an earlier saturation and larger number of failed UEs compared to NORA.

Regarding Traffic Model 2, which is depicted in figure 2.8 b), the number of succeeded UEs for ORA scheme first demonstrates a constant growth thanks to random back-off algorithm and reaches a maximum value at $k = 250$. In the meantime, the number of succeeded UEs for NORA scheme continues to rise until $k = 300$. But then they are both significantly reduced to zero when k increases from 500 to 1100 due to the excessive



(a) Traffic Model 1



(b) Traffic Model 2

Figure 2.8: The number of succeeded and failed UEs in the k-*th* RA slot of the NORA and ORA schemes under both Traffic Models[19].

collisions resulted from the accumulated failed UEs.

According to figure 2.8 the number of successful UEs for the orthogonal approach reaches its peak before and with less volume at approximately k=250 random access slots. On the other hand, at k=370 slots, the non-orthogonal scheme reaches its upper bound of 20 successful UEs. As more UEs declare access failure (a scenario portrayed by the blue curves for both OMA and NOMA) the more new UEs attempt random access all over again. As a consequence, new successful performance peaks are observed at approximately 1400-1600 for the two approaches, but we can see in the blue descending phase that the failed UEs value for the orthogonal scheme is worst (higher).

2.2.3 NOMA with CoMP

The work in [20] proposes an opportunistic NOMA scheme (ONOMA) approach for the scenario where the number of cell-edge users increases. This constitutes a problem if we apply orthogonal scheduling because the network access points allocate the same channel to a cell-edge user and this channel cannot be allocated to other users at the same time. For non-orthogonal scheduling we must bear in mind that the complexity of NOMA scales with the number of users alongside the level of multi-user interference.

The ONOMA approach has two initial phases: initialization and scheduling. In the initialization phase, B Access Points (APs) separately broadcast a normalized reference

signal *sr* to the K users, with reference transmit power P_r. Via the B reference signals, each user creates a reference power set.

In the scheduling phase each user generates a set for its preferred APs. The required CSI along with the AP selection results are fed back for K users to B APs. Then based on the feedback, the CoMP system generates B ONOMA cells.

Two problems need to be addressed in ONOMA solution: overlapping cells generated by the CoMP system, as a consequence of a user selecting more than one preferred AP, and inter-ONOMA interference, which is the interference in user k caused by APs out of the AP set S_k of that user k. Intra-ONOMA cell interference can be solved by SIC. Based on the existence or absence of overlapping cells for a user, non-ideal or ideal scenarios can be handled, respectively. For the non-ideal case, an additional algorithm is computed to overcome the overlapping cells problem.

In another work, [21] reviews the working principles of different CoMP schemes, identifying their applicability and necessary conditions for their use in a downlink multi-cell NOMA system. After that, different network scenarios with different spatial distributions of users are discussed and the applicability of CoMP schemes in these network scenarios is analyzed.

The achievable throughput of a NOMA i-*th* user and necessary condition for power allocation to perform SIC are defined, respectively, in [21], as:

$$R_i = B\log_2\left(1+\frac{\rho_i\gamma_i}{\sum_{j=i+1}^{n}\rho_j\gamma_i+1}\right), \forall i = 1,2,...,n, \tag{2.17}$$

where γ_i is the normalized channel gain with respect noise power density over NOMA bandwidth B. ρ_i is the allocated transmit power for UE-i, and must satisfy

$$\left(\rho_i - \sum_{j=i+1}^{n}\rho_j\right)\gamma_j \geq \rho_{tol}, \forall i = 1,2,...n, \tag{2.18}$$

where ρ_{tol} is the is the minimum difference in received power (normalized with respect to noise power) between the decoded signal and the non-decoded inter-user interference signals[22].

The following CoMP schemes are analyzed with respect to compatibility with NOMA approach:

- CS-CoMP (coordinated scheduling CoMP): in CS-CoMP, CoMP users are scheduled on orthogonal spectrum resources and receive desired signals only from their serving cells, respectively, while an orthogonal spectrum allocation is done based on coordination among the CoMP cells.

- CS-CoMP-NOMA: In CS-CoMP-NOMA, each CoMP user is grouped into one NOMA cluster and does not experience ICI due to orthogonal spectrum allocation among the CoMP cells.

- Jt (joint transmission)-CoMP: Jt-CoMP schemes simultaneously transmit the same data from multiple BSs to a CoMP user by using the same spectrum resources.

- Jt-CoMP-NOMA: one or more non-CoMP-users from each cell form a NOMA cluster with one or more CoMP-users. In a JT-CoMP-NOMA system, for successful decoding in presence of multiple CoMP-users in a NOMA cluster, the two following necessary conditions need to be met:

  - The signals for users receiving CoMP transmissions will be decoded prior to those for the users receiving single transmissions from their serving cells. To decode a non-CoMP user at a CoMP user equipment, the received powers for non-CoMP need to be higher than the summation of the powers of CoMP users. Although a cell can allocate more power for a non-CoMP user than the sum power of all the CoMP users in the cluster, the received power for the non-CoMP user cannot be guaranteed to be higher than the sum of received powers for CoMP users. This is because both CoMP users will receive the same signal from both the CoMP cells and thus their received powers will be improved.
  - The decoding order for a CoMP user will be the same in all NOMA clusters formed at different CoMP cells in which the CoMP user is clustered. SIC is only possible at CoMP user ends if this condition is satisfied. This condition also implies that the traditional power allocation for cell-throughput maximization will not hold in a JT-CoMP-NOMA system.

- DPS-CoMP and DPS-CoMP-NOMA: in a Dynamic Point Selection CoMP system, the data streams for each CoMP-user become available in all the CoMP-cells but only one cell sends data at a time. In each subframe, all the CoMP-cells check the channel quality for each CoMP-user, and based on the maximum channel gain only one cell is dynamically selected for data transmission. After determining the serving cell in DPS-CoMP, a CoMP-user is grouped into a NOMA cluster with the non-CoMP-users served by that cell. This scheme allows traditional power allocation like that of conventional NOMA.

- CB-CoMP: coordinated beamforming CoMP where the coordinating cells act as a distributed antenna array under a virtual BS. One CoMP-user is associated with one CoMP-cell while all the CoMP cells use same spectrum resources to serve their associated CoMP-users by utilizing the distributed MIMO principle.

To cancel ICI for CoMP-users using the same spectrum resources, the zero-forcing MIMO beamforming needs to be performed by using the CoMP-user channel vector corresponding to the CoMP-cells. Since the same beam will be used for all non-CoMP users and a CoMP-user in a CB-CoMP-NOMA cluster, the non-CoMP user may not be able to decode the message signals due to mismatch in dimension between their channel

vector (which has single dimension since only one channel exists with the serving cell) and precoding vector (which has a dimension equal to the CoMP-set size, and precoding is done based on the CoMP-users channel gains). Therefore, CB-CoMP is not applicable for a CoMP-NOMA system.

The above cited schemes can be illustrated in the figure 2.9. The presented schemes in

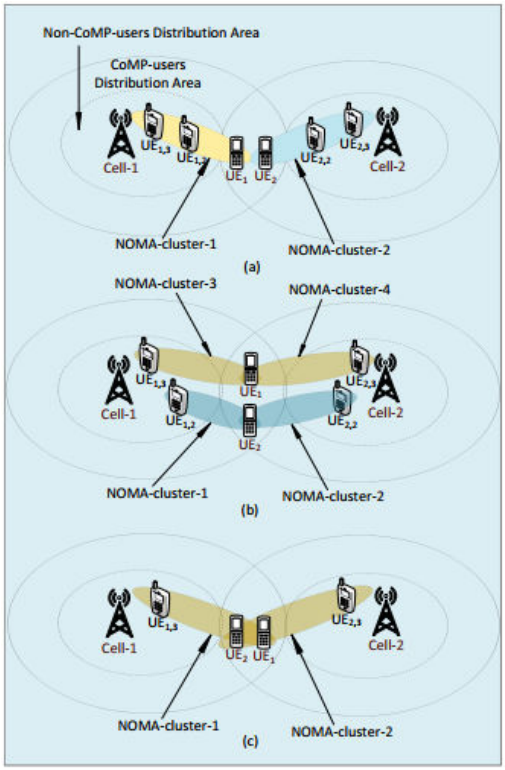


Figure 2.9: Illustrations of the various CoMP schemes for a downlink NOMA system: (a) CS-CoMP-NOMA, (b) JT-CoMP-NOMA for multiple CoMP-users and multiple non-CoMP-users, and (c) JT-CoMP-NOMA for multiple CoMP-users and a single non-CoMP-user[21]

[21] are then applied to three network scenarios and throughput equations are computed for each one of them:

Scenario 1 In this scenario, only one CoMP-user is considered for a CoMP-set, while one or multiple non-CoMP-users are considered in each CoMP-cell of that CoMP-set. By exploiting the NOMA principle, each cell superposes their NOMA users' message signals in the same spectrum resources, and thus the CoMP-user's message signal is superposed at both cells. To decode the desired signal, the decoding order for the CoMP-user needs to be same in both the NOMA clusters. Figure 2.10 represents the various CoMP-NOMA deployment scenarios: (a) deployment scenario 1, (b) deployment scenario 3:

Let $\gamma_{1,1}$, $\gamma_{1,2}$, and $\gamma_{1,3}$ denote the normalized channel power gains (with respect to noise power) for UE_1, $UE_{1,2}$ and $UE_{1,3}$ in cell 1, respectively, and $\gamma_{2,1}$, $\gamma_{2,2}$, and

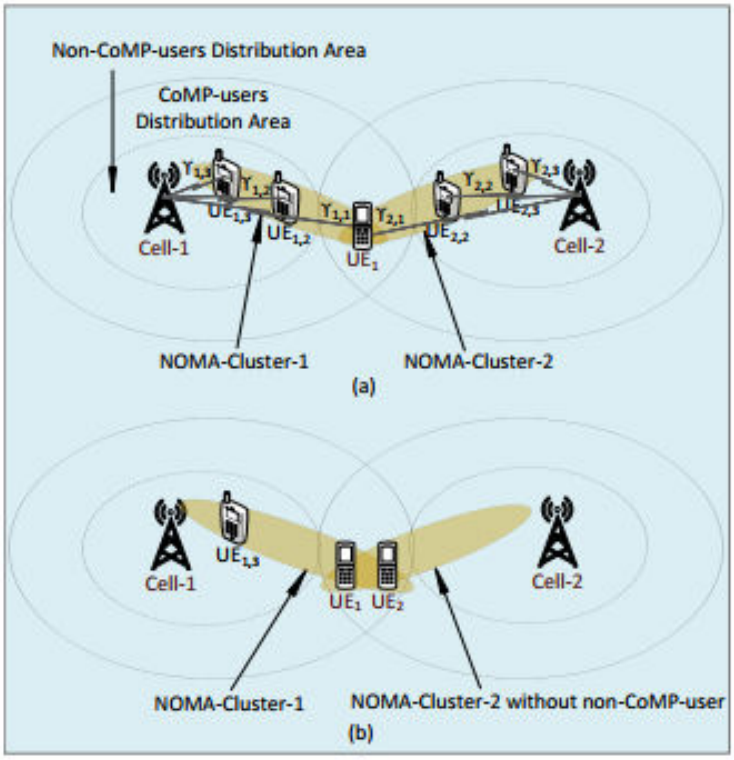


Figure 2.10: Illustrations of the various CoMP-NOMA deployment scenarios: (a) deployment scenario 1, (b) deployment scenario 3[21].

$\gamma_{2,3}$ are the normalized channel power gains for UE_1, $UE_{2,2}$ and $UE_{2,3}$ in cell 2, respectively. If the decoding order is based on the user's subscript, i.e., the message signal for UE_1 is decoded prior to decoding the other users' signals in NOMA cluster 1 and NOMA cluster 2, then the achievable throughput for the CoMP-user is:

$$R_1 = B\log_2\left(1 + \frac{\sum_{i=1}^{2}\rho_{i,1}\gamma_{i,1}}{\sum_{i=1}^{2}\sum_{j=2}^{3}\rho_{i,j}\gamma_{i,1} + 1}\right) \tag{2.19}$$

The achievable throughput for the j-*th* non-CoMP-user in cell i is:

$$R_{i,j} = B\log_2\left(1 + \frac{\rho_{i,j}\gamma_{i,j}}{\sum_{k=j+1}^{3}\rho_{i,k}\gamma_{i,j} + \sum_{m=1,m\neq i}^{2}\sum_{l=2}^{3}\rho_{m,l}\gamma'_{i,j} + 1}\right) \tag{2.20}$$

where i=1,2 and j=2,3. The term $\gamma'_{i,j}$ is the normalized channel power gain for the j-*th* non-CoMP-user in the i-*th* cell but measured from the m-*th* cell (m ≠ i) of the CoMP-set, and represents the ICI for non-CoMP-users.

If the ICI for a non-CoMP-user from any cell of a CoMP-set is negligible, then the achievable NOMA throughput for the non-CoMP-users can be approximated

as:

$$R_{i,j} = B\log_2\left(1 + \frac{\rho_{i,j}\gamma_{i,j}}{\sum_{k=j+1}^{3}\rho_{i,k}\gamma_{i,j} + 1}\right), \forall i = 1,2, \text{ and } \forall j = 2,3. \tag{2.21}$$

Scenario 2 In this scenario, [21] assumes multiple CoMP-users in a CoMP set, while one or multiple non-CoMP-users in each of the CoMP-cells of that CoMP-set. For the figure 2.9 (b), it is noted that each NOMA cluster can only include one CoMP-user, thus the spectrum resource for different CoMP-users are orthogonal. However, for the JT-CoMP-NOMA deployment scenario 2 in figure 2.9 (c), multiple CoMP-users are grouped into each NOMA cluster formed at different CoMP-cells but their decoding order will be similar in all cases. Like the scenario 1, if the decoding order is based on the user's subscript, then the achievable throughput formula for CoMP-users can be expressed as:

$$R_j = B\log_2\left(1 + \frac{\sum_{i=1}^{2}\rho_{i,j}\gamma_{i,j}}{\sum_{i=1}^{2}\sum_{k=j+1}^{3}\rho_{i,k}\gamma_{i,j} + 1}\right), \forall i = 1,2. \tag{2.22}$$

The achievable throughput for non-CoMP-users are similar to scenario 1.

Scenario 3 In a user-centric CoMP system, different CoMP-users of a particular cell can receive CoMP transmissions from CoMP-cells belonging to different CoMP-sets. In such a case, for a JT-CoMP-NOMA system, the CoMP-users of different CoMP-sets will interfere with each other and thus, will not form a NOMA cluster. Although they can form a NOMA cluster by maintaining their decoding order requirement, the ICI that [21] has neglected in scenario 1 would be excessively high. Therefore, it can be recommended that NOMA clusters are formed by including CoMP-users from one CoMP-set at a time. In this scenario, some NOMA clusters are without non-CoMP-users, and others are a mix of the two (CoMP and non-CoMP).

For spectral efficiency analysis of the referred schemes in the context of the three presented scenarios, in each NOMA cluster, the user who can decode and then cancel all the other users' signals (and hence does not experience any inter-user interference), is referred to as the cluster-head.

In all scenarios, the non-CoMP users are assumed to be at a settled distance in their spreading areas, in contrast with the random distance assigned to CoMP-users on the outskirts of the non-CoMP user's coverage areas (measured in cell-edge coverage distance). The work in [21], assumes that the NOMA cluster in one CoMP-cell is under the ascending channel power gain decoding order, as another CoMP-cell follows a different line of decoding procedure while keeping the same decoding order for CoMP-users.

In a first comparison between the Jt-CoMP-NOMA and Jt-CoMP-OMA schemes with respect to bits/s/Hz over increasing distance between BS and the cluster head in scenario 1, the simulations achieve a gain of almost two-fold in respect to the orthogonal approach. The channel gain of the link has major influence with the increasing distance.

In a comparison between Jt-CoMP-OMA and NOMA schemes and CS-CoMP-NOMA scheme for average spectral efficiency with decreasing in between cells coverage distance for the CoMP users, it is studied the advantage in spectral efficiency when adopting a JT-CoMP-NOMA approach in comparison with CS-CoMP-NOMA. As a consequence of having two CoMP-users in every JT-CoMP-NOMA cluster, the NOMA cluster that uses the best decoding order (low to high) will benefit from higher efficiency in spectrum usage than the user that employs a different order.

Du *et al*[21] also identify some challenges and issues to be dealt with for the deployment of CoMP-NOMA strategies in next generation communications. For example, an optimal power allocation for a given decoding order for each NOMA cluster is proposed in the paper. However, determining the optimal decoding order among all the coordinating cells is a challenging task. An exhaustive search algorithm could be a solution for optimal decoding order but the complexity of such a solution would be very high for a CoMP-set with more than two cells and/or two CoMP-users.

In JT-CoMP-NOMA, each CoMP-user receives the same data stream transmitted over the same spectrum resources from multiple cells, while their channel gains at each coordinating cell are different. Thus, another open research challenge is to define how much power to allocate to a JT-CoMP-user at each coordinating cell, to satisfy the user's rate requirement while achieving the optimal spectral efficiency in all the coordinating cells. On another topic, in downlink co-channel heterogenous networks, the small cell users experience strong ICI from the high power macro-cell. Since SIC is performed in the power domain, the co-channel macro-cell interference may make the small cell users unable to perform SIC. Therefore, implementation of NOMA in co-channel downlink heterogenous networks will be very challenging.

2.2.4 NOMA in a C-RAN context

In [23], eight comparisons between Orthogonal Frequency Division Multiple Access (OFDMA) and NOMA are drawn in a C-RAN environment where N BSs transmit to a central BS. The goal of [23] was to devise a power allocation scheme based on the channel gain, noise, distance and wireless propagation environment of the link connecting BS i to the central cloud BS.

In the NOMA power allocation strategy, it can be noticed in [23] that the power allocated to a BS is dependent on the power of the preceding BS having higher channel gain.

Vien *et al*[23] also computes the optimal number of BSs in the system based on total power constraints of the C-RAN and throughput performance of the cloud-edge BSs. The algorithm employed for BS number optimization stops when the rate for these cloud-edge BSs is lower than the initially defined threshold.

An OFDMA scenario, with 10 BSs with equal bandwidth and power allocation each in a C-RAN, is evaluated against a NOMA power allocation scheme with the same number of

BSs and transmission power inversely proportional to the channel gain for BS i assuming uniform noise.

It can be shown with numerical results, that total sum rate against increasing BSs number is where NOMA far surpasses OFDMA, as well as total sum-rate against increasing BSs number with different wireless propagation and fading models. NOMA outperforms OFDMA in sum-rate versus increasing SNR and marginally surpasses OFDMA in cloud-edge BSs rate with increasing BSs in the system. NOMA also allows more BSs per C-RAN with increasing cloud-edge rate, although, where NOMA gets overtaken by OFDMA is in cloud-edge rate with the latter approach obtaining better results.

2.2.5 Additional degrees of freedom in NOMA

Up until now, the literature review of this document has only shed a light and elaborated through basic-NOMA (power domain diversity) approaches and applications in the context of this thesis. In this subchapter some other diversity techniques (e.g. code, time and spatial domains) are introduced to clarify other options for NOMA. Some of the referred multiple access schemes above merely illustrate alternative degrees of freedom, which might be used throughout the development stage of this work.

2.2.5.1 SCMA: Sparse Code Multiple Access

In [24], some NOMA schemes are presented in the context of multi-carrier NOMA, which allows sub-partitions of users on the network in a single orthogonal resource block. To better understand why multi-carrier can be a reliable option we can consider the scenario where using one single carrier to group all users in a network to employ NOMA in one orthogonal resource block could be tricky, as the user with better channel state information will decode all the remaining users data before solving its own message, which leads to an increasing amount of complexity and decoding delay. In contrast, a multi-carrier NOMA approach, frequency diversity subdivides power diversity, reducing complexity since each sub-group under each carrier is limited.

Sparse Code Multiple Access (SCMA) takes advantage of a scheme that assigns a subset of the total set of carries of the system for each user. Since the cardinality of the subset of carriers is obviously smaller that the total, allied to the fact that this low spreaded feature helps to ensure that the number of users using the same subcarrier is relatively small, resulting in a somewhat manageable system complexity. The factor graph matrix is a key procedure implemented in SCMA where users are subdivided over subcarriers.

A typical factor graph matrix for a SCMA system with 6 users and 4 subcarriers can be formed like this:

$$F = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

A subcarrier i is assigned to the user j whenever $\left[F_{i,j}\right]$=1, otherwise if $\left[F_{i,j}\right]$=0, user j cannot use subcarrier i. The sparse aspect of SCMA manifests itself on the fact that each user can only use two subcarriers (each column has only two $\left[F_{i,j}\right]$=1). Multi-dimensional coding is used to guarantee that the user's data is spread over the subcarriers. Due to the fact that the data from one user at different subcarriers is jointly encoded, SCMA requires joint decoding at reception as well, which opposes to the SIC approach in the power domain NOMA schemes reception.

According to [25], from a set of J users employing SCMA, the respective encoder is set as a mapping from $\log_2 M$ binary bits to a K-dimension complex codeword chosen from j-*th* SCMA codebook with size M, where K can be seen as the spreading factor. All the K-dimensional complex codewords of the codebook are sparse vectors with $N < K$ non-zero entries. The overloading factor of the system is defined as $\lambda = \frac{J}{K}$. At reception, the received signal can be defined as:

$$y = \sum_{j=1}^{J} diag\left(\mathbf{h}_j\right)\mathbf{x}_j + \mathbf{n}, \tag{2.23}$$

where $\mathbf{x}_j = \left(x_{1,j},...,x_{K,j}\right)^T$ is the SCMA codeword of user j, $\mathbf{h}_j = \left(h_{1,j},...,h_{K,j}\right)^T$ is the channel vector of user j, and $\mathbf{n} \sim CN\left(0,\sigma^2\mathbf{I}\right)$ denotes the Gaussian noise.

Given the received signal $\mathbf{y}$ and channel knowledge $\mathbf{H} = \left(\mathbf{h}_1...,\mathbf{h}_j\right)$ at the receiver, the joint optimum MAP (maximum *a posteriori*) detection will estimate $\widehat{\mathbf{X}}$ that maximizes the joint a posteriori probability (APP) mass function of the transmitted codeword $\mathbf{X} = \left(\mathbf{x}_1...,\mathbf{x}_j\right)$, i.e., $p\left(\mathbf{X}|y\right)$, and can be given by:

$$\widehat{X} = \arg\max p\left(\mathbf{X}|\mathbf{y}\right), \mathbf{X} \in \left(x_{j=1}^{J}\right)\chi_j, \tag{2.24}$$

where $\left(x_{j=1}^{J}\right)\chi_j = \chi_1,...,\chi_J,\chi_j$ is the codeword set of user j. Unfortunately, the necessary cost for exactly computing the optimal MAP detector increases exponentially with respect to the number of users J, which limits its application in practical systems. In [25], a shuffled-MPA multiuser detection scheme is proposed to improve receiver complexity.

2.2.5.2 Network Diversity Multiple Access (NDMA)

In Network Diversity Multiple Access (NDMA), received packets that have collided are stored in memory rather than being discarded. They are later combined with future retransmissions to extract all the collided information packets. The BS forces terminals to transmit P copies of each packet when P terminals collide [26]. The technique exploits

diversity combining ideas to distinguish the conjoined packets. The traditional diversity methods are created using multiple antennas for reception, however, such is not the case for NDMA where network resources are employed to assure diversity through a careful selection of retransmissions[27]. But P copies might not suffice in weak propagation conditions.

Ganhão *et al*[26] suggests a Hybrid-ARQ Network Diversity Multiple Access (H-NDMA) approach which, instead of asking for P copies of each packet every time that P packets collide, uses time-diversity for slotted random access where the access mechanism forces the UEs involved in a collision with reception errors, to retransmit more than P times.

In H-NDMA the uplink slots are organized in a sequence of epochs. The BS broadcasts a control SYNC packet through the assigned downlink channel, thus, announcing the beginning of a new epoch and allowing any UE with data packets to transmit, to do so in the next slot. The UEs that do not transmit in the first epoch are not allowed to transmit until the next SYNC. The BS can discern all colliding packets using specific orthogonal identification sequences for each terminal. In the first slot of an epoch, the BS detects collisions and uses a downlink channel to advertise them, asking all the involved UEs to retransmit. When P>1 UEs are involved in a collision, the BS asks for P-1 retransmissions for P packets separation. After this initial P-slots set, the BS recognizes the data packets reception and can ask for up to R additional retransmissions using H-ARQ[9], for the packets that were received with errors.

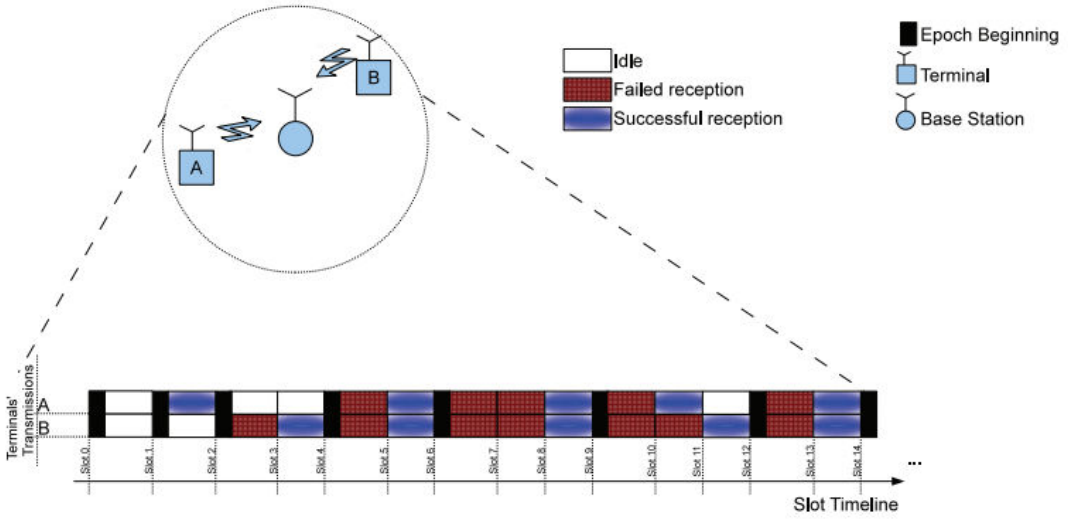


Figure 2.11: H-ARQ multipacket reception scheme [26].

Diversity Combining (DC) and CC are the two techniques employed in error correction in [28]. When employing a CC scheme, a number of copies of a packer are used to

[9]Hybrid-ARQ (Automatic Repeat reQuest) constitutes a set of techniques where error detection (ED) and forward error correction (FEC) are combined to mitigate interference.

form noise-corrupted codewords with growing codewords length and smaller rate codes. In the DC approach, the individual symbols from equal copies of a data packet are linked to form a singular packet with more steady symbols.

The scheme proposed in [28] allows for different retransmission techniques that require rearrangement of the data block before transmission.

Two extremes cases concerning the channel conditions are considered in the paper. In Uncorrelated Channel (UC) the channel is uncorrelated in between more attempts to retransmit. In other case, known as Equal Channel (EC), channel conditions remain permanent for all retransmission efforts. If, on the other hand, UC condition is assumed, all packet retransmissions from each UE are uncorrelated, meaning that the channel response changes completely for each retranmission l of a given UE p. Such assumption of having little to no correlations between transmissions for DC and MPR schemes might be ideal since it is not doable in the majority of network practical scenarios, as long as the time interval between retransmission remains short, or the transmission frequency does not change significantly between attempts. Therefore, the EC condition displays an under performance comparing it with UC, but a much more realistic one. In order to avoid an ill-conditioned matrix inversion for MPR, small correlations under EC conditions are employed. The work in [28] assumes two different retransmission schemes: Equal Channel with Phase Rotation (ECPR) and Equal Channel with Shifted Packet (SP).

With the ECPR technique, different phase rotations are employed for the transmitted data blocks by each UE at different retransmission attempts. Assuming that θ_p is a fixed phase for UE p and δ_l is an offset for each transmission then a phase rotation $\left(\theta_p + \delta_l\right)^l$ is applied for the l-*th* retransmission of the p-*th* UE data block, which is formally equivalent to have $H_{k,p}^{(l)} = H_{k,p}^{(1)} \exp\left(j(\theta_p + \delta_l)\right)$. By performing the phase rotations, the matrix inversions required for MPR are well conditioned. However, since the ECPR does not change the magnitude of $H_{k,p}^{(l)}$, it is useless for DC.

With the SP technique a cyclic shift ζ_l is performed on the frequency-domain samples associated to the l-*th* retransmitted block. For example, if a UE attempts to retransmit a block, it performs a cyclic shift of $\zeta_2 = \mathrm{N}/2$; if the reception fails, it retransmits with a cyclic shift $\zeta_3 = \mathrm{N}/4$, etc. This is formally equivalent to perform a cyclic shift to the channel response of a given UE, i.e., $H_{k,p}^{(l)} = H_{(k+\zeta_l) mod N,p}^{(1)}$. This is especially efficient with time-dispersive channels, where the frequency response changes substantially from subcarrier to subcarrier. This technique can be interesting not just to avoid ill-conditioned matrix inversions in MPR but also to avoid deep in-band fades in DC[28].

In the context of MTC-related approaches that adequate to this specific kind of traffic, Ramos *et al* [29] proposed a new random-access Medium Access Control (MAC) protocol for MTC systems, which applies MPR techniques to provide energy efficient interaction and extreme low latency to MTC UEs.

For the next chapter, the author tries to apply spatial and power diversity over H-NDMA in order to calculate the computational toll that such diversity schemes impose on

network requirements, and study how these strategies can improve latency and reliability in a cloud computing context.

Chapter 3

URLLC Implementation using IB-DFE in a C-RAN

This chapter presents the simulations and tests made to investigate which degrees of freedom in a radio network system can be used to reduce the terminals processing and access time considering an IB-DFE receiver at the BSs deployed over a given topology distribution of UEs. The goal is to fulfill, or at least approach, URLLC requirements. Manipulating the diversity strategies through power, time or space dimensions, it is possible to configure the network in order to decide where processing operations can be held, either locally or centrally in a cloud, to help minimize or scale latency and reliability requirements.

The algorithm used in the simulations in this chapter takes advantage of the H-NDMA protocol concept (introduced in 2.2.5.2) where, instead of demanding a new copy of a packet everytime the reception from a given UE fails, H-NDMA combines the failed packets received to increase the reliability of the system. Therefore, what is shown in this chapter is the variation in copies and computing time needed under different diversity conditions. These results are obtained considering also more realistic channel response models.

3.1 C-RAN Architecture

The architecture considered is depicted in figure 3.1. The terminals closer to the RRH are solved locally within the cell, and the ones on the cell-edge must be carefully managed through the correct usage of space, power and time dimensions. Also, the deployment of slicing algorithms at the virtual BBU pool is needed to ensure the best possible performance. URLLC requires a very high reliability and low latency (LL). High reliability was imposed by enforcing a maximum measured PER value below a threshold ρ [1] for a given UE_p belonging to a set S UEs being solved.

Low latency performance measures two components: the medium access time and the receiving algorithm processing time. Two metrics are considered for each of these components: the number of retransmissions (a measurement of the radio access time, consisting on the number of needed copies of a packet that the UE with the maximum PER in a subset of S solved UEs has to send to achieve a PER of ρ) and the computation time, which is just the time from the start to finish of the resolution of a transmission with L packets at the IB-DFE receivers used in this work.

The main goal of this chapter is aimed at observing the output of various C-RAN configurations comprising the degrees of freedom used in a radio network system and in a C-RAN cloud, and try to reach an SINR estimation value at the receiver to guarantee LL reliability in the uplink channel. On top of said metric, an algorithm can be built to allocate network processing load and attain for URLLC requirements.

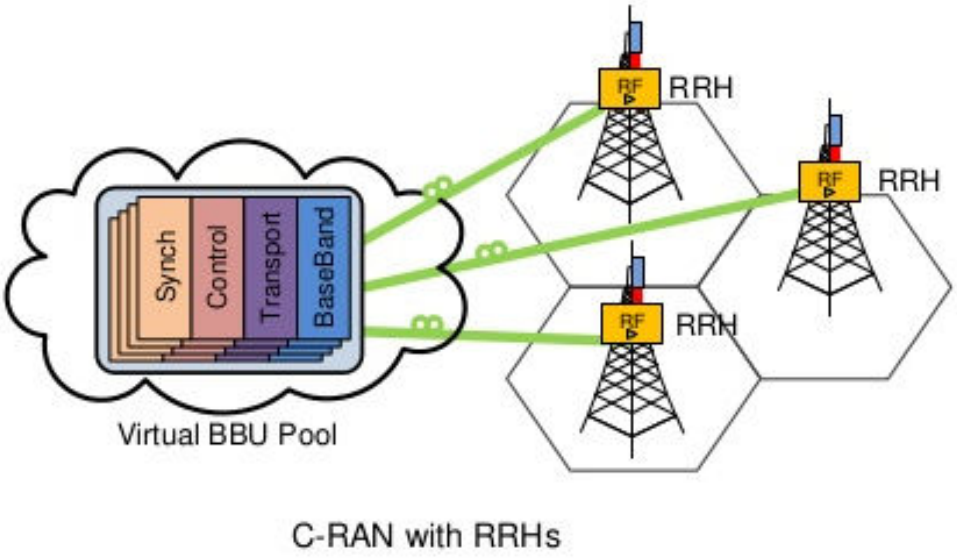


Figure 3.1: Architecture model for C-RAN 5G communication

[1] ρ is the maximum allowed PER in this thesis set to 10^{-3}.The packet error ratio PER is the number of incorrectly received data packets divided by the total number of received packets in the simulations.

3.2 Uplink Diversity techniques

Throughout this chapter, whenever space and time freedom degrees are employed, we may be talking of CoMP, which consists in multiple BSs combining their signals in a logical BBU, or Hybrid CoMP, which is just the same concept as CoMP with the additional time dimension, where signals are combined in a matrix of the form:

$$\begin{bmatrix} Y_k^{(1)} \\ \vdots \\ Y_k^{(L)} \end{bmatrix} = \begin{bmatrix} \left|\xi_{1,1}\right| H_{k,1}^{(1)} & \dots & \left|\xi_{1,P}\right| H_{k,P}^{(1)} \\ \vdots & \ddots & \vdots \\ \left|\xi_{L,1}\right| H_{k,1}^{(L)} & \dots & \left|\xi_{L,P}\right| H_{k,P}^{(L)} \end{bmatrix} \begin{bmatrix} S_{k,1} \\ \vdots \\ S_{k,P} \end{bmatrix} + \begin{bmatrix} N_k^{(1)} \\ \vdots \\ N_k^{(L)} \end{bmatrix}, \tag{3.1}$$

where each line of the matrix represents a spatial degree of freedom (a BS) and $\left|\xi_{L,1}\right|$ is the attenuation from BS L to UE. Hybrid CoMP merely adds a time freedom degree to this matrix, i.e. if P UEs need to retransmit 3 times to every BS in the system, the matrix would have 3 (retransmissions) x L (BSs) lines. This matrix structure helps to diminish interference because it utilizes all the signals from one UE to all BSs which in the partial case wouldn't happen since only one signal is solved and the other 2 (in a 3 BS scenario) are viewed as interference (see figure 3.2). Equation (3.1) can consider all the signals transmitted, or a partial resolution, where some may be excluded, and handled as noise. Power NOMA (or P-NOMA) is an example where partial resolution is applied.

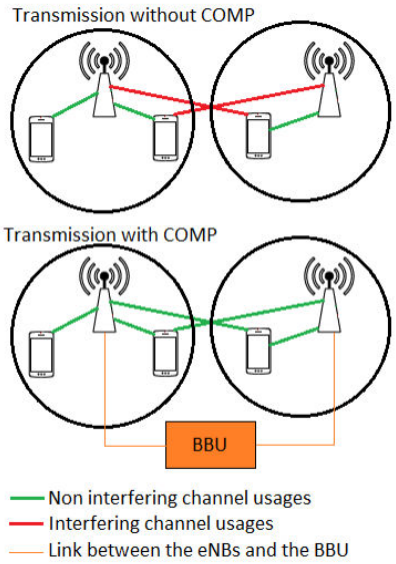


Figure 3.2: Partial model against CoMP model scheme

Besides managing the correct approach towards choosing the fastest way to solve UEs throughout the topology just by deciding to aggregate signals in a joint BBU or solving them locally, another degree of freedom is also studied which pertains to the arriving signal power of the UEs to the BSs. In a concept known as power-NOMA, which is based in power gaps between arriving signals at a BS, diversity can be useful in slicing traffic in the network. This signal power can be handled in order to achieve optimal separation at the IB-DFE receiver to perform SIC, or even to reduce the receiving algorithm processing overhead, allowing the distribution over multiple BBUs.

3.3 IB-DFE Receiver

An IB-DFE receiver is employed to deal with diversity schemes in an integrated fashion. In the dissertation, one of the main subjects is to investigate the suitability of such receiver in providing URLLC services. This section presents the mathematical specification of the receiver for the three diversity modes, generalizing [28] model. In [28], a PER is calculated. On top of this PER which is set to a threshold for the simulations (usually ρ), the diversity schemes can be evaluated in terms of number of retransmissions and computation time. The IB-DFE receiver decodes the UE's L transmissions up to N iterations. The estimated data symbol $\tilde{S}_{k,p}^{(i)}$ for a given iteration I and UE p is:

$$\tilde{S}_{k,p}^{(i)} = F_{k,p}^{(i)^T} Y_k - B_{k,p}^{(i)^T} S_k^{(i-1)}, \tag{3.2}$$

where $F_{k,p}^{(i)^T} = \left[F_{k,p}^{(i,1)},...,F_{k,p}^{(i,L)}\right]$ are the feedforward coefficients and $B_{k,p}^{(i)^T} = \left[B_{k,p}^{(i,1)},...,B_{k,p}^{(i,P)}\right]$ are the feedback coefficients. $\bar{S}_k^{(i-1)} = \left[\bar{S}_{k,1}^{(i-1)},...,\bar{S}_{k,p}^{(i-1)}\right]^T$ are the soft decision estimates from the previous iteration for all UEs. $\bar{S}_k^{(i-1)}$ can be related to the symbol's hard decisions, $\hat{S}_k^{(i-1)}$, where according to [30] results

$$\bar{S}_k^{(i-1)} \simeq P^{(i-1)} \hat{S}_k^{(i-1)}, \tag{3.3a}$$

$$\hat{S}_k^{(i-1)} = P^{(i-1)} + \Delta_k, \tag{3.3b}$$

$P^{(i-1)} = diag(\rho_1^{i-1},...,\rho_P^{i-1})$ are the correlation coefficients and $\Delta_k = \left[\Delta_{k,1},...,\Delta_{k,P}\right]^T$ is a zero mean error vector. Expanding for a given UE p, assuming a QPSK constellation, where according to [31] results

$$\rho_P^{(i-1)} = \frac{1}{2N} \sum_{n=0}^{N-1} \left|\rho_{n,p}^{I^{(i-1)}}\right| + \left|\rho_{n,p}^{Q^{(i-1)}}\right|, \tag{3.4}$$

so that

$$\rho_{n,p}^{I^{(i-1)}} = \tanh\left(\frac{\left|L_{n,p}^{I^{(i-1)}}\right|}{2}\right), \tag{3.5a}$$

$$\rho_{n,p}^{Q^{(i-1)}} = \tanh\left(\frac{\left|L_{n,p}^{Q^{(i-1)}}\right|}{2}\right), \tag{3.5b}$$

where

$$L_{n,p}^{I^{(i-1)}} = \frac{2}{\sigma_{n,p}^{2^{(i-1)}}} Re\left\{\tilde{s}_{n,p}^{(i-1)}\right\}, \tag{3.6a}$$

$$L_{n,p}^{Q^{(i-1)}} = \frac{2}{\sigma_{n,p}^{2^{(i-1)}}} Im\left\{\tilde{s}_{n,p}^{(i-1)}\right\}, \tag{3.6b}$$

and

$$\sigma_{n,p}^{2^{(i-1)}} = \frac{1}{2N}\sum_{n'=0}^{N-1}\left|\hat{s}_{n',p}^{(i-1)} - s_{n',p}\right|^2. \tag{3.7}$$

For the first iteration, i.e. i=1, $\tilde{S}_k^{(i-1)}$ is a null vector and $P^{(i-1)}$ is a null matrix. Assuming that R_S, R_N and R_Δ, are respectively, the correlation of S_k, N_k and Δ_k, where

$$R_S = \mathbb{E}\left[S_k S_k^H\right] = 2\sigma_S^2 I_P, \tag{3.8a}$$

$$R_N = \mathbb{E}\left[N_k N_k^H\right] = 2\sigma_N^2 I_L, \tag{3.8b}$$

$$R_\Delta = \mathbb{E}\left[\Delta_k \Delta_k^H\right] \simeq 2\sigma_S^2\left(I_P - P^{(i-1)^2}\right). \tag{3.8c}$$

σ_S^2 is the symbol's variance, I_P is the identity matrix with length P and σ_N^2 is the noise's variance. Knowing that $\Gamma_p = \left[\Gamma_{p,1}=0,...,\Gamma_{p,p}=1,...,\Gamma_{p,P}=0\right]^T$ and

$$\alpha_{k,p}^{(i)} = F_{k,p}^{(i)} H_k^T - B_{k,p}^{(i)} P^{(i-1)^2} - \Gamma_p, \tag{3.9a}$$

$$\beta_{k,p}^{(i)} = B_{k,p}^{(i)^T} P^{(i-1)}, \tag{3.9b}$$

the Mean Square Error (MSE), $\mathbb{E}\left[\left|S_{k,p} - \tilde{S}_{k,p}^{(i)}\right|^2\right]$, of $S_{k,p}$ is

$$\mathbb{E}\left[\left|S_{k,p} - \tilde{S}_{k,p}^{(i)}\right|^2\right] = \mathbb{E}\left[\left|\Xi_{k,p}^{(i)} - \Upsilon_{k,p}^{(i)}\right|^2\right], \tag{3.10}$$

where $\Xi_{k,p}^{(i)} = F_{k,p}^{(i)^T}\left(H_k^T S_k + N_k\right)$ and $\Upsilon_{k,p} = B_{k,p}^{(i)^T} P^{(i-1)}\left(P^{(i-1)} S_k + \Delta_k\right) + \Gamma_p S_k$. Expanding (3.10) results

$$\begin{aligned}\mathbb{E}\left[\left|S_{k,p}-\tilde{S}_{k,p}^{(i)}\right|^2\right]=\\ \mathbb{E}\left[\left|\left(F_{k,p}^{(i)^T}H_k^T-B_{k,p}^{(i)^T}P^{(i-1)^2}-\Gamma_p\right)S_k\right|^2\right]...\\ +\mathbb{E}\left[\left|B_{k,p}^{(i)^T}P^{(i)}\Delta_k\right|^2\right]+\mathbb{E}\left[\left|F_{k,p}^{(i)^T}N_k\right|^2\right]=\\ \alpha_{k,p}^{(i)^*}R_S\alpha_{k,p}^{(i)^T}+F_{k,p}^{(i)^H}R_NF_{k,p}^{(i)}+\beta_{k,p}^{(i)^*}R_\Delta\beta_{k,p}^{(i)^T}.\end{aligned} \tag{3.11}$$

To obtain the optimal coefficients, $F_{k,p}^{(i)}$ and $B_{k,p}^{(i)}$, under the minimum Mean Square Error (MSE) criterion, the gradient of the Lagrange function is applied to (3.11). So

$$\nabla J=\nabla\left(\mathbb{E}\left[\left|S_{k,p}-\tilde{S}_{k,p}^{(i)}\right|^2\right]+\left(\gamma_p^{(i)}-1\right)\lambda_p^{(i)}\right), \tag{3.12}$$

where the Lagrange multipliers are constrained to $\gamma_p^{(i)}-1=\frac{1}{N}\sum_{k=0}^{N-1}\sum_{l=1}^{L}F_{k,p}^{(i,l)}H_{k,p}^{(l)}-1$.

From the following set of equations,

$$\begin{cases}\nabla_{F_{k,p}^{(i)}}J=0\\ \nabla_{B_{k,p}^{(i)}}J=0\\ \nabla_{\lambda_p^{(i)}}J=0\end{cases}, \tag{3.13}$$

$\nabla_{F_{k,p}^{(i)}}J=0$ is

$$\begin{aligned}H_k^HR_SH_kF_{k,p}^{(i)}-H_k^HR_SP^{(i-1)^2}B_{k,p}^{(i)}-H_k^HR_S...\\ +R_NF_{k,p}^{(i)}+\frac{1}{N}H_k^H\lambda_p^{(i)}\Gamma_p=0,\end{aligned} \tag{3.14}$$

$\nabla_{B_{k,p}^{(i)}}J=0$ is

$$\left(P^{(i-1)^2}R_S+R_\Delta\right)B_{k,p}^{(i)}=R_SH_kF_{k,p}^{(i)}-R_S\Gamma_p, \tag{3.15}$$

and $\nabla_{\lambda_p^{(i)}}J=0$ when $\gamma_p^{(i)}=1$. So the optimal coefficients are

$$\begin{cases}B_{k,p}^{(i)}=H_kF_{k,p}^{(i)}-\Gamma_p\\ F_{k,p}^{(i)}=\Lambda_{k,p}^{(i)}H_k^H-\Theta_{k,p}^i\end{cases}, \tag{3.16}$$

$\Lambda_{k,p}^{(i)}=\left(H_k^H\left(I_P-P^{(i-1)^2}\right)H_k+\frac{\sigma_N^2}{\sigma_S^2}I_L\right)^{-1}$ and $\Theta_{k,p}^i=\left(I_P-P^{(i-1)^2}\right)\Gamma_p-\frac{\lambda_p^{(i)}}{2\sigma_S^2N}\Gamma_p$. For a single UE p

transmitting data, i.e. without collisions, results

$$\gamma_p^{(i)} = 1, \tag{3.17a}$$

$$B_{k,p}^{(i)} = \sum_{l=1}^{L} F_{k,p}^{(i,l)} H_{k,p}^{(l)} - 1, \tag{3.17b}$$

$$F_{k,p}^{(i,l)} = \frac{\gamma_p^{(i)} H_{k,p}^{l^*}}{\frac{\sigma_N^2}{\sigma_S^2} + \sum_{l=1}^{L} \left| H_{k,p}^{(l)} \right|^2}. \tag{3.17c}$$

From equation (3.11), and the optimal $F_{k,p}^{i}$ and $B_{k,p}^{i}$ coefficients from equation (3.16), it is possible to compute the minimum MSE. Considering that

$$\sigma_p^{2^{(i)}} = \frac{1}{N^2} \sum_{k=0}^{N-1} \mathbb{E}\left[\left| \tilde{S}_{k,p}^{(i)} - S_{k,p} \right|^2 \right], \tag{3.18}$$

and $Q(x)$ as the Gaussian error function, then in accordance with [32], the Bit Error Rate (BER) of UEs p at the ith iteration for a QPSK constellation is

$$BER_p^{(i)} \simeq Q\left(\frac{1}{\sigma_p^{(i)}} \right). \tag{3.19}$$

Although the linear frequency-domain receiver can be employed with any constellation, the presented iterative receiver is specific for QPSK constellations. The iterative receiver can be extended to other constellations by employing the generalized IB-DFE receiver concept of [33]. For M^2-QAM constellations with Gray mapping and minimum distance between symbols (i.e., the real and imaginary parts of the symbols are $\pm 1, \pm 3, \ldots,$ $\pm$(M-1)), the BER is approximately given by

$$BER_p^{(i)} \simeq \frac{2}{\log_2(M)} \left(1 - \frac{1}{M} \right) Q\left(\frac{1}{\sigma_p^{(i)}} \right). \tag{3.20}$$

For an uncoded system with independent and isolated errors, the PER for a fixed packet size of M bits is

$$PER_p^{(i)} \simeq 1 - \left(1 - BER_p^{(i)} \right)^M. \tag{3.21}$$

To show how PER varies with the number of retransmissions from an UE to an IB-DFE receiver, the work in [28], that employs the same receiver model of this thesis, illustrates in figure 3.3 the gain in PER observed using different L_s (copies) of the same packet. It is clearly visible that the values in blue have a earlier drop in PER with respect to normalized power E_b/N_0 needed at the receiver. Which means that with more copies of the same packet, for the same number of iterations in the IB-DFE receiver, less power can be employed in transmission. It can also be seen the gain obtained with four iterations compared to the linear MMSE receiver, with one iteration. This particular characteristic

of the IB-DFE receiver, which is less power demanding than other linear receivers, might suit the high power saving requirements of MTC traffic.

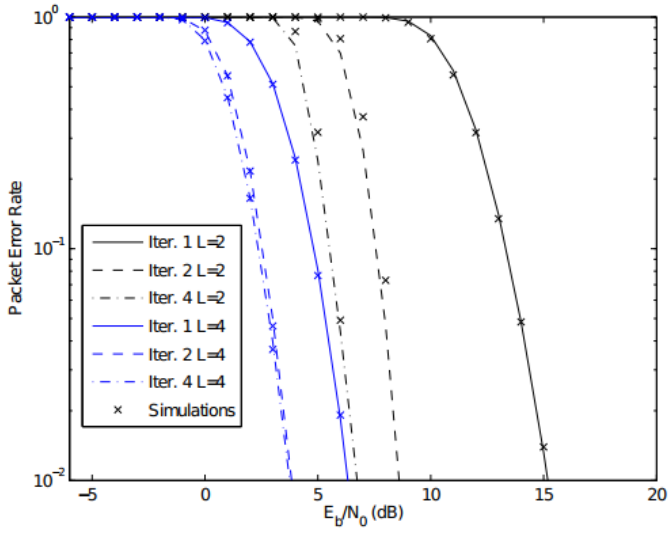


Figure 3.3: Multipacket Detection for P = 2 UEs, L = [2, 4] and up to 4 iterations regarding an iterative receiver structure.

3.4 Channel Diversity: Uncorrelated channel and Shifted packet scenarios

In an uncorrelated channel response scenario, all the results are obtained under the assumption that between transmissions, the channel response for a given UE changes completely, which is not a practical premise to use if we want to model a real mobile network system. A channel cannot be completely uncorrelated along such short time spans (between transmission *n* and transmission *n-1* for the same UE). In this section a solution to handle fixed channels is analyzed. Following [34], for systems where uncorrelated channel is not practical, we could assume that the frequency domain block associated to the rs^{th} retransmission of the p^{th} packet, $\left\{A_{k,p}^{(r)}; k = 0,1,...,N-1\right\}$, is an interleaved version of $\left\{A_{k,p}; k = 0,1,...,N-1\right\}$ (using time-domain interleaving is not an option since it increases significantly the complexity of the packet separation). Since this is formally equivalent to assume that $\left\{H_{k,p}^{(r)}; k = 0,1,...,N-1\right\}\left(r = 2,...,N_p\right)$ is an interleaved version of $\left\{H_{k,p}; k = 0,1,...,N-1\right\}$, the channel correlations for each frequency can be very small. However, to avoid transmitting signals with very large envelope fluctuations, it is better to assume that $\left\{A_{k,p}^{(r)} = A_{k+\zeta_{r,p}}; k = 0,1,...,N-1\right\}$, i.e., it is a cyclic-shifted version of $\left\{A_{k,p}; k = 0,1,...,N-1\right\}$, with shift ζ_r . This means that the corresponding time-domain block is $\left\{a_{n,p}^{(r)} = a_{n,p}e^{\frac{j2\pi\zeta_r n}{N}}; n = 0,1,...,N-1\right\}$, with a suitable ζ_r. Therefore, this technique

is formally equivalent to have $A_{k,p}^{(r)} = A_{k,p}$ and $H_{k,p}^{(r)}$ a cyclic-shifted version of $H_{k,p}^{(1)}$, with shift $-\zeta_r$.

In general, the larger ζ_r the smaller the correlation between $H_{k,p}^{(r)}$ and $H_{k,p}^{(1)}$, provided that $\zeta_r < N/2$ (since we consider cyclic shifts, $\zeta_r = N$ is equivalent to have $\zeta_r = 0$). In [34], as well on this dissertation it is assumed that the different ζ_r are the odd multiples of N/2, N/4, N/8, etc., i.e.,

$$\begin{aligned}
\zeta_2 &= \frac{N}{2}, \\
\zeta_3 &= \frac{N}{4}, \\
\zeta_4 &= \frac{3N}{4}, \\
\zeta_5 &= \frac{N}{8}, \\
\zeta_6 &= \frac{3N}{8}, \\
\zeta_7 &= \frac{5N}{8}, \\
\zeta_8 &= \frac{7N}{8} \ldots
\end{aligned}$$

3.5 The interference model

The IB-DFE receiver model includes the effect of interfering UEs, i.e, UEs that while being served by a given BS b, impact on every other BS in the system due to their transmission power, and thus the power from terminals that are not served by a given BS b must be taken into consideration in our model. Assuming the approximation that the multipath effects can be despised for the interfering UEs, the interference for an UE p is approximated by $\left|\xi_p\right| S_{k,p}$. The interference on transmission k is given by:

$$\phi_k \approx \sum_{p=P+1}^{M} \left|\xi_p\right| S_{k,p}, \tag{3.22}$$

where M stands for the total number og UEs considered. The expanded expression for Y_k is:

$$\begin{bmatrix} Y_k^{(1)} \\ \vdots \\ Y_k^{(L)} \end{bmatrix} = \begin{bmatrix} |\xi_1| H_{k,1}^{(1)} & \ldots & |\xi_1| H_{k,P}^{(1)} \\ \vdots & \ddots & \vdots \\ |\xi_1| H_{k,1}^{(L)} & \ldots & |\xi_P| H_{k,P}^{(L)} \end{bmatrix} \begin{bmatrix} S_{k,1} \\ \vdots \\ S_{k,P} \end{bmatrix} + \begin{bmatrix} N_k^{(1)} + \phi_k^{(1)} \\ \vdots \\ N_k^{(L)} + \phi_k^{(L)} \end{bmatrix}. \tag{3.23}$$

In the presence of spatial diversity, different values of ϕ_k should be considered for each retransmission because interference is felt differently in each BS. Both interference and noise are modeled as Normal distribution functions so the sum of these functions is also a Normal distribution given by:

$$\mathbb{E}\Big[N_k\Big] = \mathbb{E}\Big[\phi_k\Big] = 0, \tag{3.24}$$

and

$$\mathbb{E}\Big[N_k^{eq}\Big] = \mathbb{E}\Big[N_k + \phi_k\Big] = 0. \tag{3.25}$$

It is assumed that the interference and noise are two independent Normal distributed random variables. Interference can vary between the different transmissions, so its variance can take different values. For a system where L copies of the signal are received, the interference variance is $\varsigma_\phi^2 = diag\Big(\sigma_\phi^{2^{(1)}}, ..., \sigma_\phi^{2^{(L)}}\Big)$.Therefore the variance of N_k^{eq} for a retransmission i is:

$$\sigma_{N_k^{eq}}^{2^{(i)}} = \sigma_{N_k}^{2^{(i)}} + \sigma_{\phi_k}^{2^{(i)}}, \tag{3.26}$$

and

$$\varsigma_{N_k^{eq}} = \varsigma_{N_k} + \varsigma_{\phi_k}. \tag{3.27}$$

The noise correlation matrix used in the equation of the MMSE for the estimation of $\tilde{S}_{k,p}^{(i)}$ is given by:

$$\mathbb{E}\Big[N_k^{eq} N_k^{eq^H}\Big] = \mathbb{E}\Big[\Big(N_k + \phi_k\Big)\Big(N_k + \phi_k\Big)^H\Big] = \mathbb{E}\Big[\Big(N_k N_k^H\Big)\Big(\phi_k \phi_k^H\Big)\Big]. \tag{3.28}$$

The optimal feed forward coefficients obtained under the MMSE criteria is now given by:

$$F_{k,p}^{(i)} = \Lambda_{k,p}^{(i)} H_k^H \Theta_{k,p}^{(i)}, \tag{3.29}$$

where $\Lambda_{k,p}^{(i)} = \Big(H_k^H\Big(I_P - P^{(i-1)^2}\Big)H_k + \frac{1}{\sigma_S^2}\varsigma_{N_k^{eq}}^2\Big)^{-1}$ and $\Theta_{k,p}^{(i)} = \Big(I_P - P^{(i-1)^2}\Big)\Gamma_p - \frac{\lambda_p^{(i)}}{2\sigma_S^2 N}\Gamma_p$.

3.6 Simulation deployment and results

This section provides the research content and results of the implementation of said diversity strategies throughout two types of environments: one and two cell (BSs) topologies. Besides, a study is provided beforehand with respect to the reliability results of the hardware where the research was conducted before presenting the computing time and access time (L copies of a packet for each subset of UEs being solved) results. The goal is to reach a compromise between access and computing time to alleviate workload between network nodes.

3.6.1 Topology distribution

To conduct the diversity tests shown throughout the chapter, two topologies were devised and UEs transmission powers were adjusted accordingly to best emulate a radio network scenario. The topologies comprise two fentocell scenarios: one with a 100 m^2 area with

two BSs and another with a 25 m^2 area with only one BS. The spreading of mobile terminals through the two areas is based on a 2D Poisson distribution shown in the following expression:

$$P(X = c) = \frac{\left(\rho_{ON}\beta A_E\right)^c}{c!} e^{-\rho_{ON}\beta A_E}, c = 0, 1, ..., N. \tag{3.30}$$

This Poisson distribution reflects the number of users in an area of the network, represented by the random variable X. ρ_{ON} represents the probability of finding a device transmitting, which is set to one on both figures, A_E is just the area considered for the probability distribution (100 m^2 and 25 m^2) and N is the maximum number of UEs. After obtaining the number of UEs transmitting in the squared area, their coordinates are obtained using a uniformly distributed pseudo-random number generator, to assure a random spacing between them. The BSs coordinates are predefined with a $4m$ distance between the two for figure 3.5.

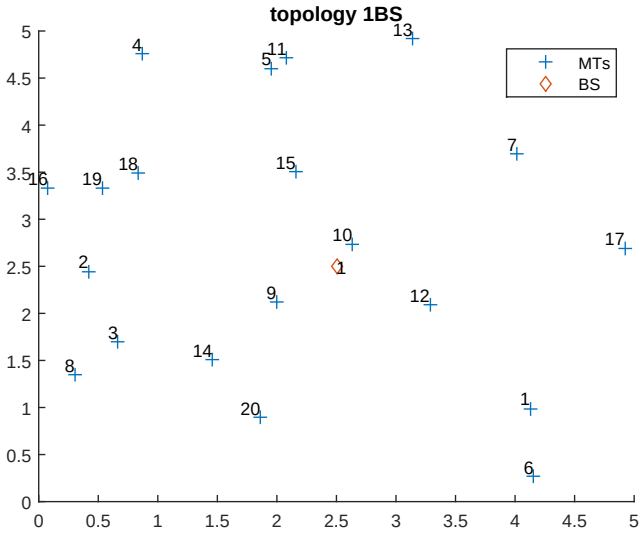


Figure 3.4: Topology deployed for diversity simulations with one BS

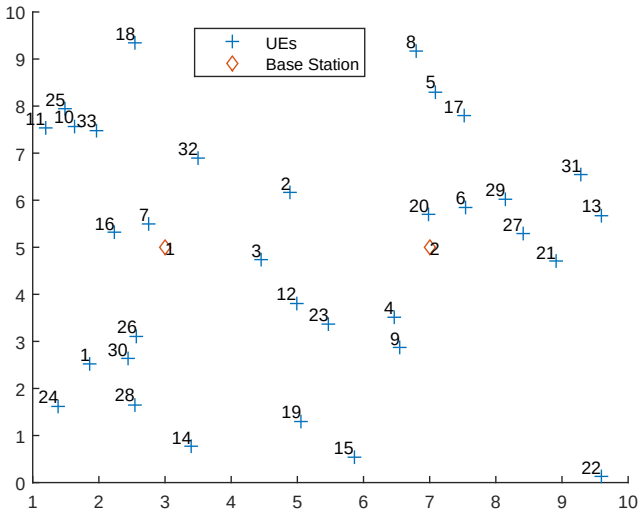


Figure 3.5: Topology deployed for diversity simulations with two BSs

3.6.1.1 UE transmission power calculation

The UEs in figure 3.5 are first associated to each BS by their distance to each one of the two. If $\Delta_{UE_p \longrightarrow BS_1} = min(\Delta_{UE_p \longrightarrow BS_j}, \forall j)$ then UE_p associates (is solved) itself to BS_1 - which then becomes the BS_{Near} - with

$$\Delta_{UE_p \longrightarrow BS_1} = \sqrt{\left(x_{UE_p} - x_{BS_1}\right)^2 + \left(y_{UE_p} - y_{BS_1}\right)^2}. \tag{3.31}$$

The path loss is calculated with respect to the BS that UE_p belongs to, but the path loss regarding the other *j-1* BSs is computed to study interference of UE_p on the set of the remaining *j-1* BSs. In this case, the number of j is equal to two, with *far* denominating the farthest BS from UE_p. Thus, two path losses are computed: $PL_{Near} = -10n\log_{10}\left(\Delta_{UE_p \longrightarrow BS_{Near}}\right)$ and $PL_{Far} = -10n\log_{10}\left(\Delta_{UE_p \longrightarrow BS_{Far}}\right)$, where n is the path loss coefficient and $\Delta_{UE_p \longrightarrow BS_{\{Near,Far\}}}$ are the distances between the receivers and the transmitter. Transmission power P_{txNear} for $\Delta_{UE_p} \longrightarrow BS_{Near}$ is then calculated with respect to the normalized power at the BS_{Near} (EN_{BSNear}, or E_b/N_0 at BS_{Near}), which is just the power expected at the receiver and the value used for normalization in reception of all the power values, this meaning that if there is need for the UEs served by a given BS to raise their transmission power, the value of EN_{BSNear} can be increased, thus

$$P_{txNear} = |PL_{Near}| + EN_{BSNear} dBs. \tag{3.32}$$

The $\left|\xi_p\right|$ attenuation coefficient is just the result of 3.32 converted to linear scale, so

$$\left|\xi_p\right| = \sqrt{10^{\left(\frac{P_{txNear}}{10}\right)}}, \tag{3.33}$$

and

$$\phi_{p,BS_{Far}} = \sqrt{10^{\left(\frac{P_{txFar}}{10}\right)}} S_{k,p}, \tag{3.34}$$

which is the interference of UE_p on BS_{Far}. LTE standards do not let a UE transmit over 3dBs (or 33dBm). In this dissertation we defined the same constraint for the transmission power, P_{tx}. We refer to the following constraint:

$$P_{txRealNear} = |PL_{Near}| + P_{Reception} \leq 3dB, \tag{3.35}$$

where

$$P_{Reception} = E_{N_{BSNear}} + \sigma^2_{N_0} + G_0 + 10\log_{10} B, \tag{3.36}$$

being $\sigma^2_{N_0} = -174 + 10\log_{10} H$ the thermal noise for a bandwidth H. B denotes the number of bits in a packet (B=2 in the simulations considered), and G_0 is the antenna gain. For UEs in figure 3.4 the transmission power calculation follows only equation (3.32), without the need for pre-association from the UEs to the nearest BS, since there is only one BS to receive the entire set of UEs. Equation (3.32) can be generally used for any given EN_{BSNear} at the reception.

3.6.2 Normalization of values between cores

In the processing performance shown in the figures throughout this chapter, the data curves were obtained by simulation in different machines. Overall there are 3 processing families where the code was run: AMD Ryzen 7 (core 3), Intel Core i7 (core 4) and Intel Xeon (core 5). The specifications for each core can be seen in table 3.1.

Core ID	Model Name	RAM Size
4	Intel(R) Core(TM) i7-4790S CPU @ 3.20GHz	8 GB
5	Intel(R) Xeon(R) CPU ES-1620 v2 @ 3.70GHz	64 GB
3	AMD Ryzen 7 1700 Eight-Core Processor	16 GB

Table 3.1: Specs for all three simulation machines used

The number of copies L needed for a number of N UEs to reach an overall maximum PER of ρ can vary, as shown in figure 3.3 with the E_b/N_0 (normalized power) and also with the number of iterations at the IB-DFE receiver. Although it might not vary between machines, the time each machine takes to solve a matrix of 128 bits x L copies x N UEs might serve as a performance indicator of the differences between three distinct hardware simulation environments. The difference in performance between the three cores is analyzed in a simulation of partial resolution where each UE belonging to a subset of S (cardinality of subset S is equal to value in x axis, for a given computation time $y(x)$, being left out in resolution, for each value of x, the total number of UEs minus the cardinality of subset of the strongest S UEs in the topology) UEs increase retransmissions of a packet until the PER of the UE_p - with UE_p belonging to the subset of current UEs being solved, S - with the highest PER value reaches ρ. Along the y axis, computation time at the receiver is observed against increasing number of UEs being solved. No interference was considered. The topology employed is displayed in figure 3.5, and figures 3.6 to 3.14 represent the concatenated data for both BSs, and three types of fitting are represented for each core: second degree polynomial fitting, first degree polynomial fitting and finally, power fitting.

The values with respect to the goodness of each fitting, and curve coeficients were obtained for both BSs performance on the three cores (tables 3.3 and 3.4). These metrics help to justify irregularities on some core curves, even for such small population of UEs envolved.

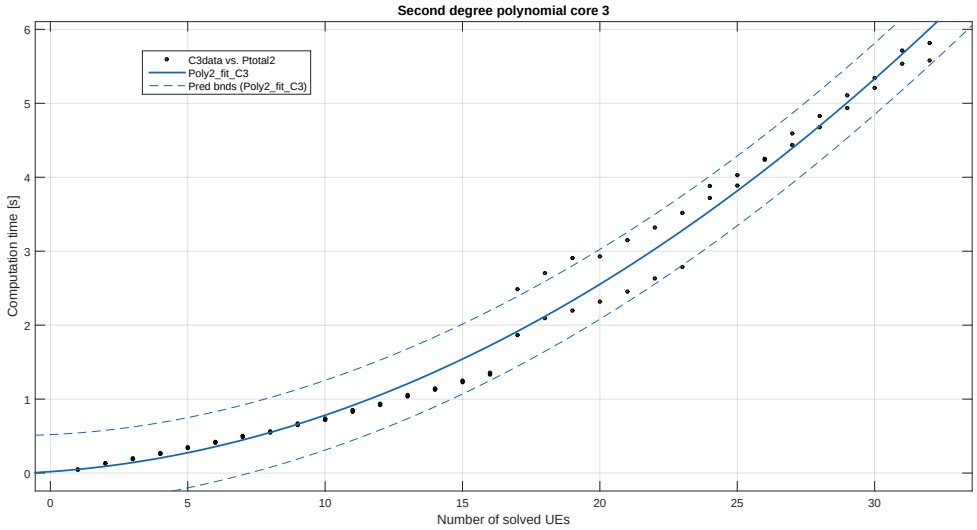


Figure 3.6: Computation time for core 3 with second degree polynomial fitting curve from concatenated data obtained employing partial solving at both BSs for the topology in figure 3.5

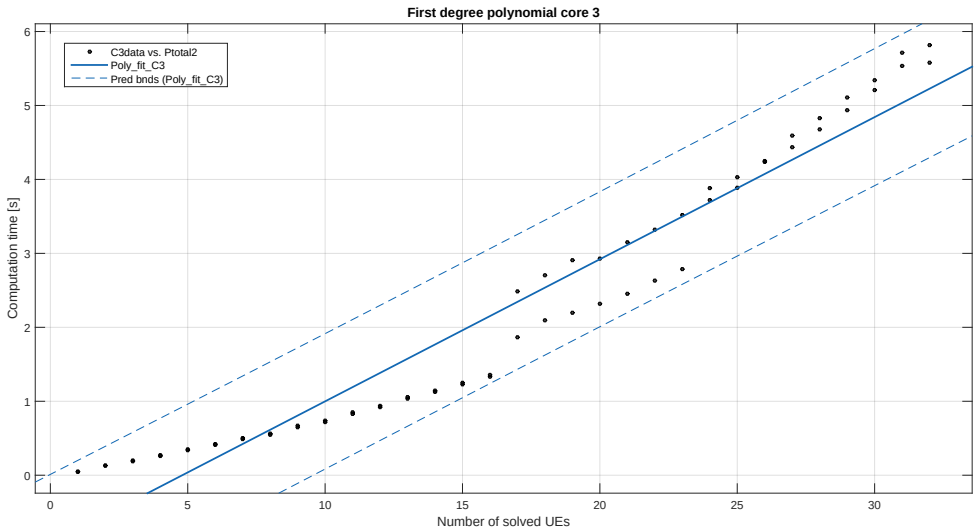


Figure 3.7: Computation time for core 3 with first degree polynomial fitting curve from concatenated data obtained employing partial solving at both BSs for the topology in figure 3.5

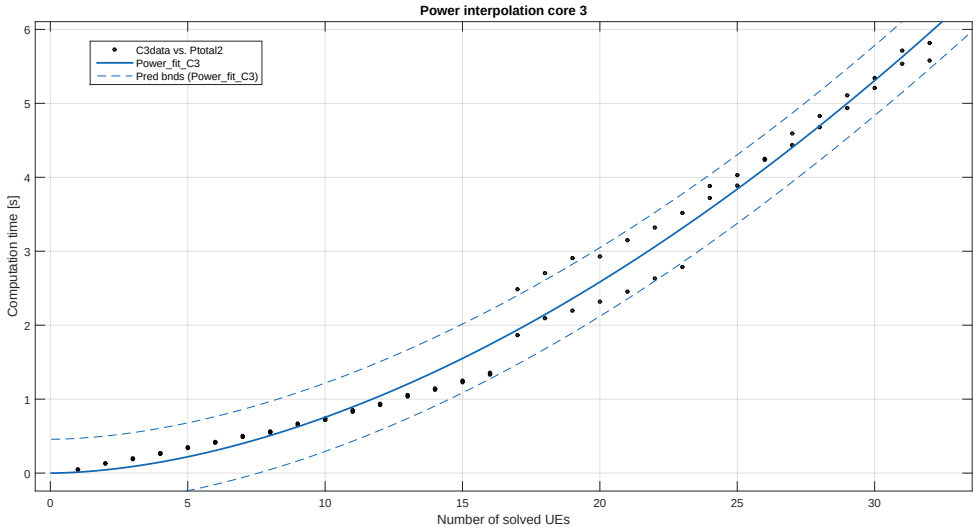


Figure 3.8: Computation time for core 3 with power fitting curve from concatenated data obtained employing partial solving at both BSs for the topology in figure 3.5

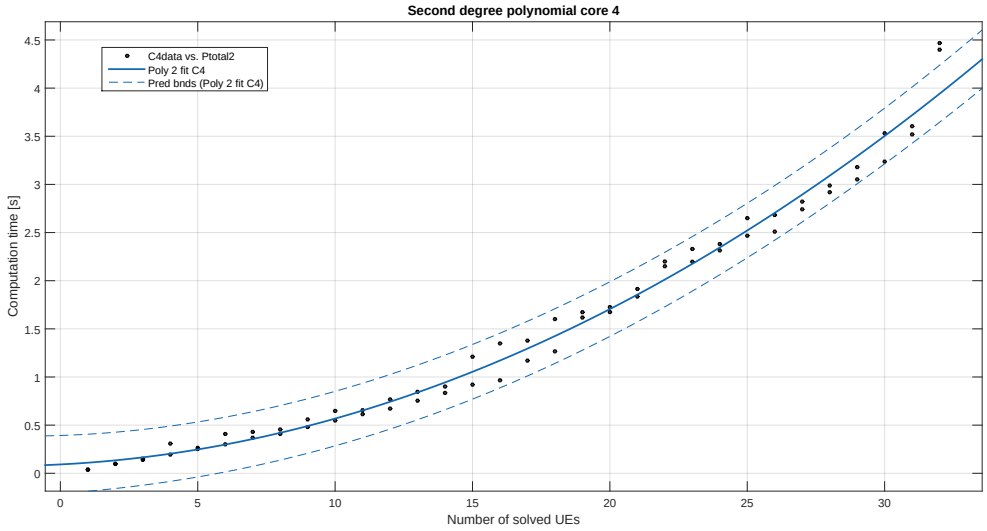


Figure 3.9: Computation time for core 4 with second degree polynomial fitting curve from concatenated data obtained employing partial solving at both BSs for the topology in figure 3.5

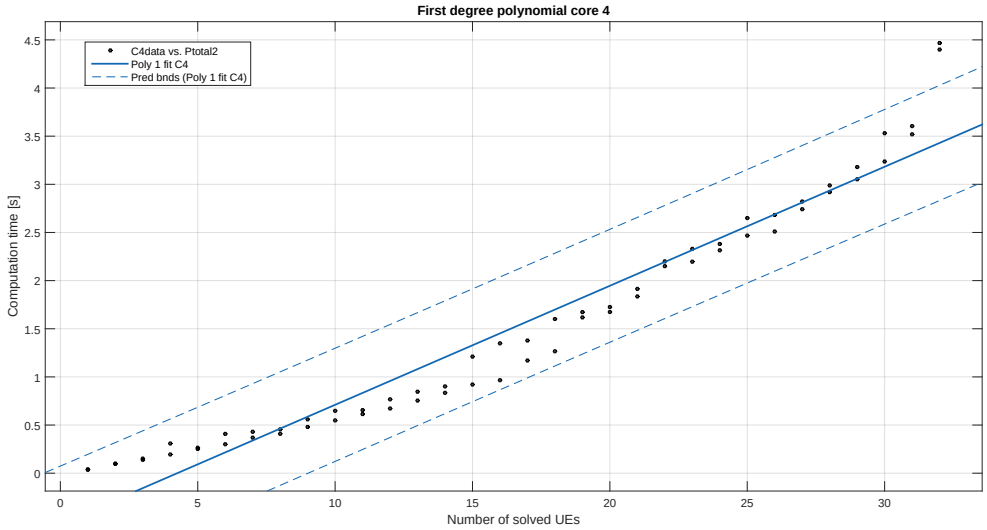


Figure 3.10: Computation time for core 4 with first degree polynomial fitting curve from concatenated data obtained employing partial solving at both BSs for the topology in figure 3.5

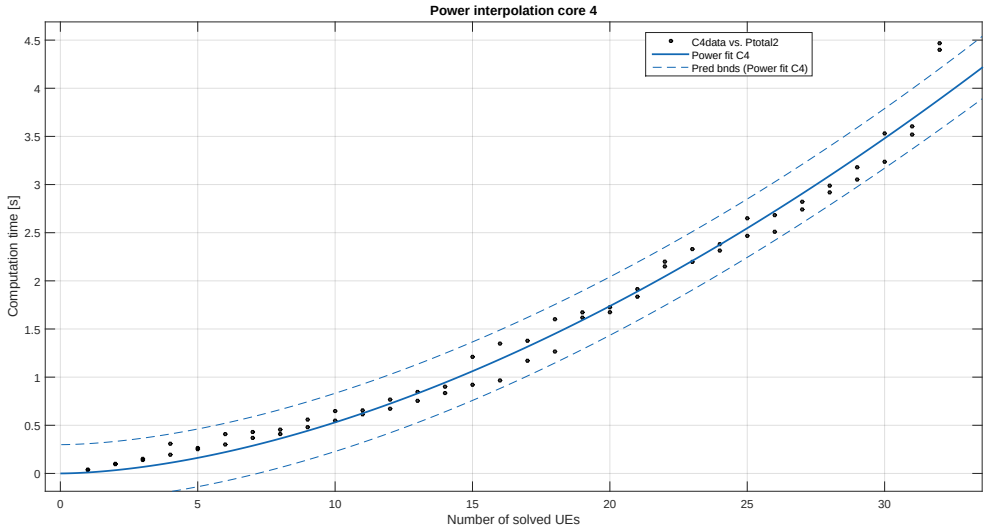


Figure 3.11: Computation time for core 4 with power fitting curve from concatenated data obtained employing partial solving at both BSs for the topology in figure 3.5

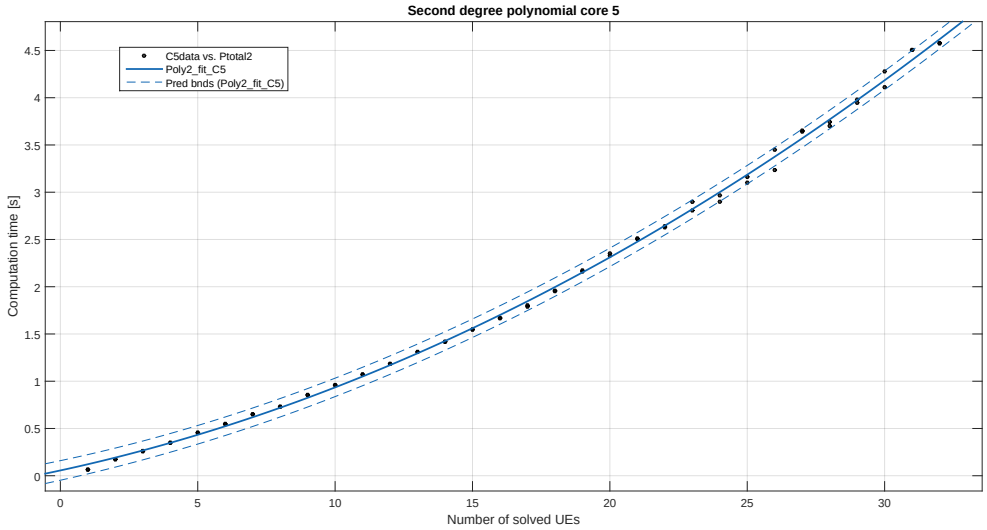


Figure 3.12: Computation time for core 5 with second degree polynomial fitting curve from concatenated data obtained employing partial solving at both BSs for the topology in figure 3.5

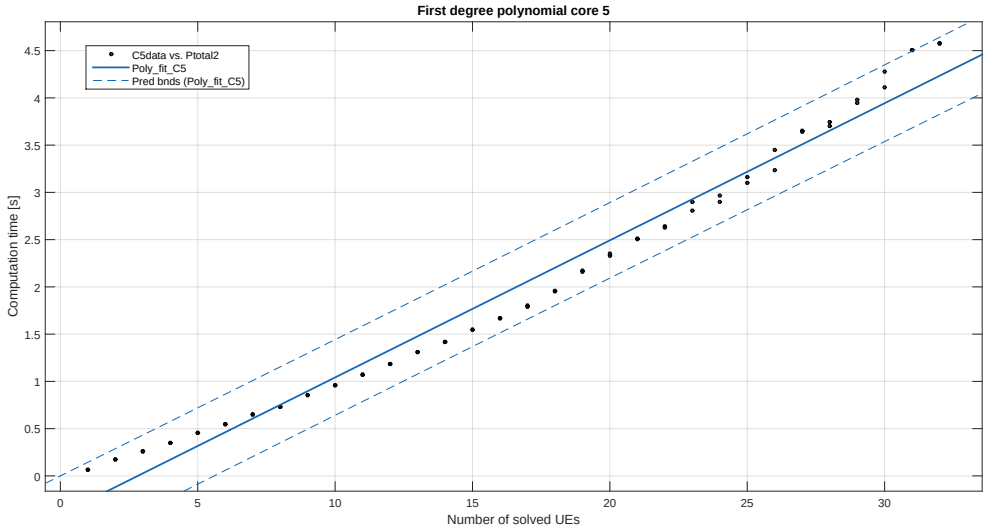


Figure 3.13: Computation time for core 5 with first degree polynomial fitting curve from concatenated data obtained employing partial solving at both BSs for the topology in figure 3.5

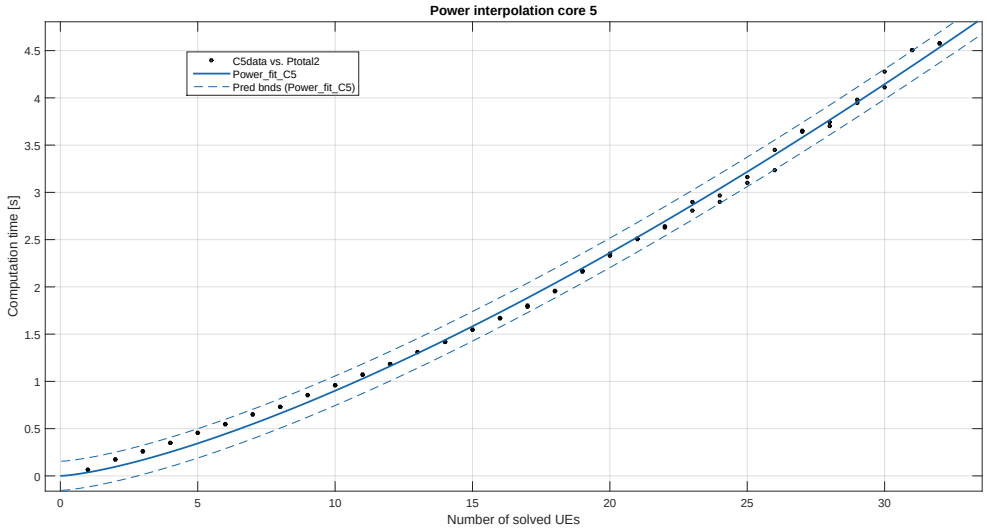


Figure 3.14: Computation time for core 5 with power fitting curve from concatenated data obtained employing partial solving at both BSs for the topology in figure 3.5

Power fitting curve	Core ID		
Goodness of fit	3	4	5
SSE	3.2456	1.3853	0.3682
R-Square	0.9848	0.9844	0.9969
RMSE	0.2288	0.1495	0.0771

Table 3.2: Power fitting curve statistics for all three cores

First degree polynomial fitting curve	Core ID		
Goodness of fit	3	4	5
SSE	12.7062	5.2391	2.4338
R-Square	0.9407	0.9409	0.9792
RMSE	0.4527	0.2907	0.1981

Table 3.3: First degree polynomial fitting curve statistics for all three cores

Second degree polynomial fitting curve	Core ID		
Goodness of fit	3	4	5
SSE	3.2944	1.1886	0.1412
R-Square	0.9846	0.9866	0.9988
RMSE	0.2324	0.1396	0.0481

Table 3.4: Second degree polynomial fitting curve statistics for all three cores

Figures 3.6 to 3.14 depict the computing time used to run the H-NDMA receiver algorithm for the three machines. Tables 3.2 to 3.4 show the 4 metrics results employed for each of the three fits. Since the data samples are relatively scarce volume wise, it is much more meaningful to try to predict a suitable expression for each machine behaviour. SSE (Sum of Squares Due to Error) measures the total deviation of the response values from the fit to the response values, a value closer to 0 indicates that the model has a smaller random error component, and that the fit will be more useful for prediction. R-Square values measures how successful the fit is in explaining the variation of the data. Put another way, R-square is the square of the correlation between the response values and the predicted response values. R-square can take any value between 0 and 1, with a value closer to 1 indicating that a greater proportion of variance is accounted for by the model. For example, an R-square value of 0.8234 means that the fit explains 82.34% of the total variation in the data about the average. Root Mean Squared Error (RMSE) is also known as the fit standard error and the standard error of the regression. It is an estimate of the standard deviation of the random component in the data. Under the said metrics, we can see that core 5 acomplishes the best predictor in comparison with the other two cores, and the second degree polynomial is the best tailored fitting curve to achieve such result.

It is visible that due to the lower CPU frequency of the AMD Ryzen 7, core 3 has a less stable performance: intermediate for a low number of users, where concurrent multi-core processing seems to compensate it. But for more than 17 UEs, the performance degrades

and becomes the worst. The best performance was achieved with Core 4, with only 8GB of RAM, due to the small memory footprint of the process. But for more than 30 UEs, the performance degrades. It should be emphasized that no GPU processing was used, which would allow a significant reduction on the processing times. In order to satisfy an end-to-end URLLC requirement, we need to keep the H-NDMA matrices dimensions low to contain the computational time. Techniques are needed to split complex matrices in different concurrent computations.

3.6.3 Diversity simulations

Adopting the context of the topologies in figures 3.4 and 3.5, diversity schemes like CoMP and power-NOMA used for radio network systems are employed in the following subsections to search for a way to subdivide processing load in the network, thus reaching new strategies for latency and reliability improvement. The output observed results from the experiment of increasing copies of a packet per user in order to reach a PER of ρ. Once this ρ is achieved the experiment proceeds in increasing the number of users solved, as interferents decrease as a consequence. Also it is relevant for the scope of the thesis the evaluation of the SINR for the weakest UE being solved in each experiment, because the weakest UE power signal at the receiver can be a trustworthy indicator of how reliable is the communication and how efficient the power distribution between UEs in the topology is being done. Therefore, the author searches for correlation in performance peaks of SINR relatively to the same peaks in computation time and retransmission count. This SINR analysis is further extended in section 3.6.3.4.

3.6.3.1 Hybrid CoMP vs Partial solving

Following the model described in section 3.2, a comparison was drawn between partial solving, i.e, only one BS solving an entire matrix of 128 bits x L copies x N UEs, and a Hybrid CoMP strategy where both BSs in the figure 3.5 cooperate and combine their received signals in a matrix of the form 128 bits x L*number of BSs x N UEs. L is our unknown variable at the start of the experiment, and is selected to satisfy the reliability requirement. We already know that in figure 3.5 N is 33.

In this test, L(i.e the number copies per UE) is increased until the matrix PER satisfies the threshold ρ (following the expression in (3.21)). The solving time values in the table express the total computation time to solve a matrix of 128 bits x L copies of a packet per user x 33 users, divided by the number of code runs for each datapoint obtained, which in this experiment equals to 1000 times.

The values of L and the matrix resolution times are observed in table 3.4. We can see that a gain in packet copies per user can be attained with Hybrid-CoMP, but, for the topology in figure 3.5, there is a penalising effect in solving time. We can perhaps choose to combine and switch between partial and coordinated multipoint stategies if we want to process UEs that reach BSs with similar power, because those UEs will be the ones

Strategy	Average Solving Time [s]	Matrix Size [bits x copies x users]
Partial BS1	3.826	128x21x33
Partial BS2	3.158	128x22x33
CoMP BBU	4.666	128x11x33

Table 3.5: Performance measurements between the partial strategy in both BSs and Hybrid-CoMP

that most benefit from the combination of signals in a joint BBU due to similar distance between BSs resulting in high arriving power to both.

3.6.3.2 Channel diversity studies

As shown in section 3.4, an uncorrelated channel response system is unlikely to occur in practical radio network scenarios. Therefore, in this section tests are deployed in order to compare an ideal uncorrelated scenario with a fixed shifted packet (SP) channel transmission. The tests shown in the figures (3.15) and (3.16) translate the experiment of increasing the number of copies for n UEs being solved in a matrix, until the overall PER reaches the nominal goal of ρ. As more UEs are considered in the matrix, more copies of the same packet are needed, as expected. No interference is presented just for benchmarking: it calculates the performance when no interference arrives to neighbour BS, i.e. the arriving power from UE p to BS_{far} is equal to zero. No power NOMA scheme is applied. The results were simulated in core 3 for the topology in figure 3.5.

As expected, in figures 3.15 and 3.16, there is degradation in terms of needed copies to reach a PER of $\rho = 10^{-3}$ when the channel is not fully uncorrelated, meaning that the use of SP, despite being well suited for our scenario in the sense that it is more realistic, subtracts somewhat of a percentage of the channel diversity from the system, in comparison with a fully uncorrelated channel response assumption.

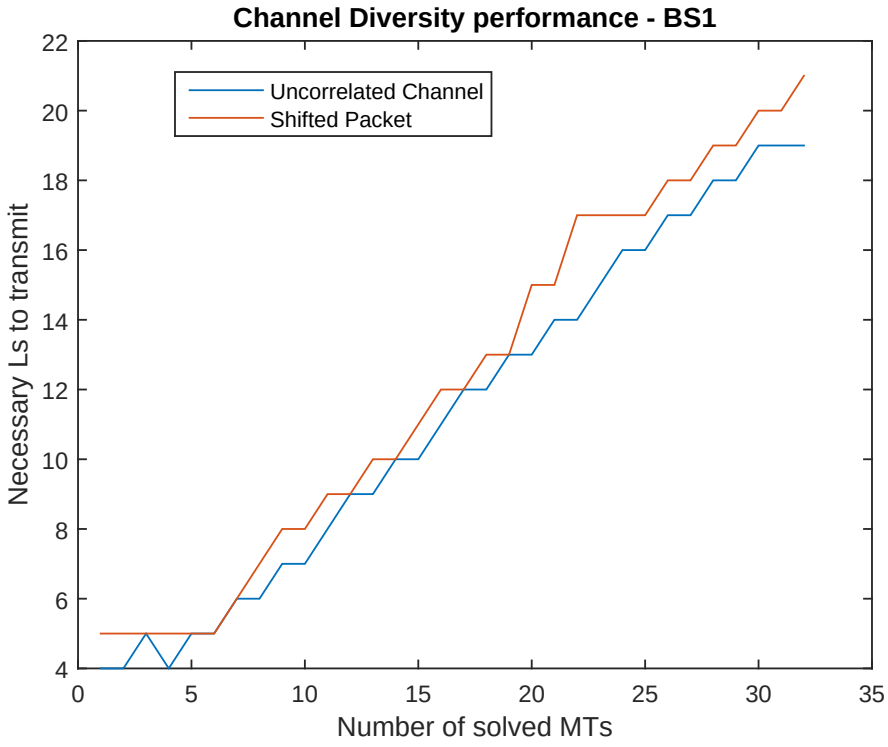


Figure 3.15: Difference in retransmission demands for uncorrelated and shifted packet channel conditions for BS1

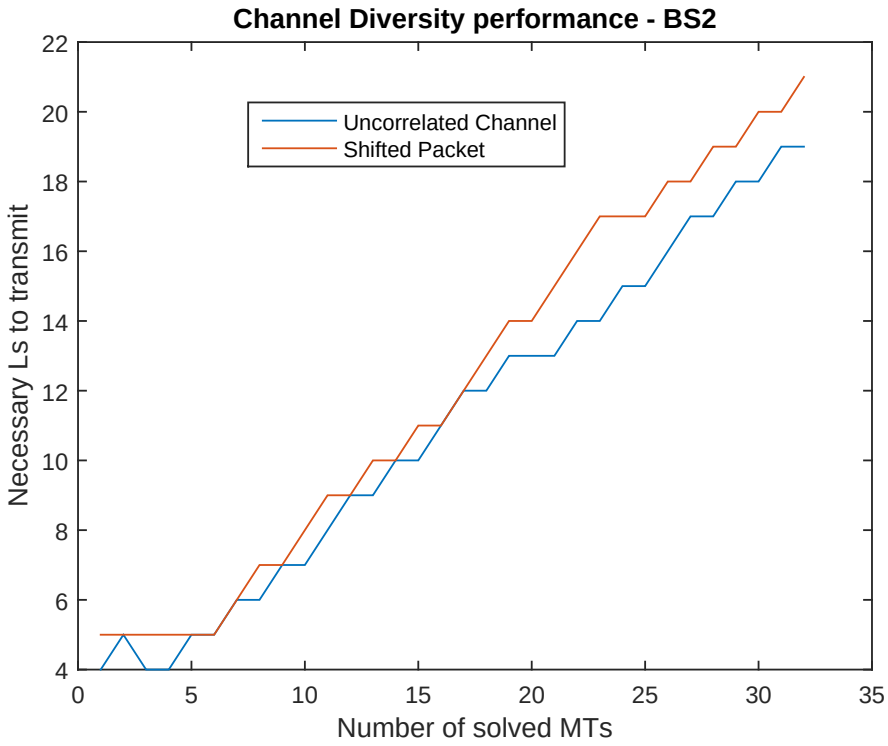


Figure 3.16: Difference in retransmission demands for uncorrelated and shifted packet channel conditions for BS2

3.6.3.3 Power diversity studies

As mentioned in section 3.2, besides time and space dimensions, power diversity at the uplink can also be explored and examined to tune the best configurations in order to decrease delay in the network. In this section, the author shows studies that exemplify how transmission power allocation can be a factor to consider in the network model to attain URLLC requirements. The two topologies of figures 3.4 and 3.5 are used to exemplify the discrepancies in performance when the power diversity schemes commute from one to a two-cell scenario.

One BS scenario

In order to observe direct benefits of using power multiplexing at the receiver that results from the power diversity at the uplink, for the topology in figure 3.4, a SIC scenario was devised where for 30 different power gaps the $P/2$ nearest UEs, where P is the total number of UEs in figure 3.4, transmit with a given number of copies until the UE with the highest PER in $P/2$ high power UEs equals ρ. Then, the second strongest set of $P/2$ UEs is solved.

On top of the uplink power diversity, three types of channel conditions are also depicted simultaneously: the already examined uncorrelated and shifted packet (SP) scenarios, and a third type of channel, AWGN (additive white gaussian noise, with a merely illustrative purpose in figure 3.17 since it is not a realistic channel model for wireless scenarios), where the absolute value of the channel response coefficients between copies is always equal to one. For the latter channel, due to the constant absolute value between copies which eliminates the channel diversity potential, a poorer performance is expected.

The output of the experiment in terms of copies is observed in figure 3.17. The chosen normalized power E_b/N_0 at the receiver for this experiment was 20 dBs, simply because the stretch of 30 dBs over the minimum required value expected at the reception extracted from expression (3.35) still allowed the author to remain under the LTE limit for transmission power. Thus the author was still under the restraints of the following expression:

$$P_{txRealNear} = |PL_{Near}| + P_{Reception} + PowerGap \leq 3dB, \tag{3.37}$$

with a power gap set ranging from 0 to 30 dBs, that the author incremented for a total of 30 dBs tested in figure 3.17.

For time efficiency reasons, a maximum value of Ls=90 transmissions per MT (a simulation sealing) is imposed for the epoch durations in this test. Whenever this limit is reached and one or more UEs PER is above ρ, the configuration is not valid - points with Ls equal to 90 in the figures represent invalid configurations. In figure 3.17 it is visible that, when the gap reaches 15dBs between highest power level and second highest power level for two sets of UEs (nearest and farthest), the test for one BS assumes a tail behavior almost

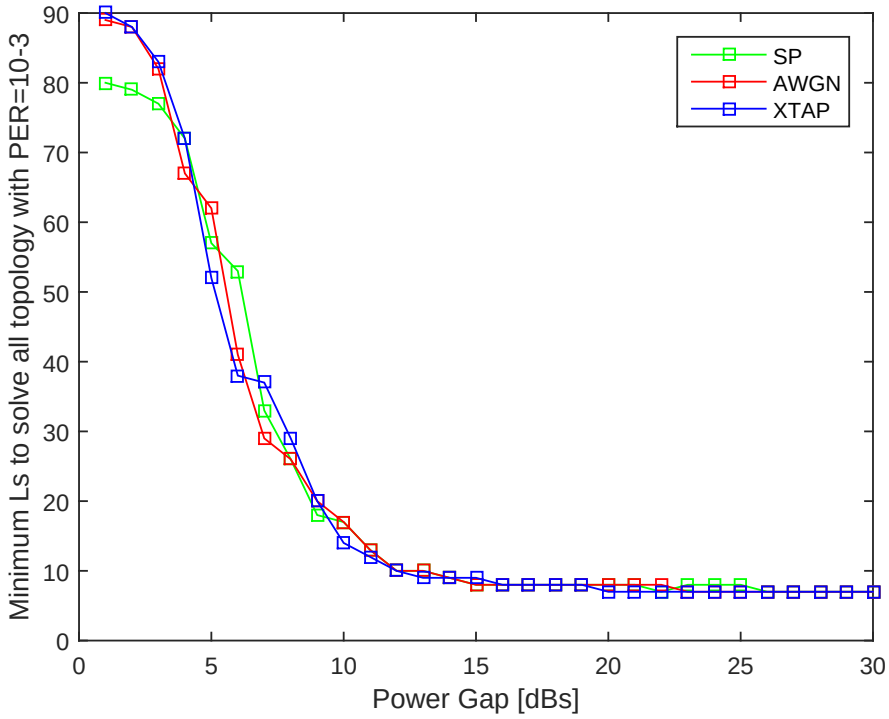


Figure 3.17: Necessary amount of Ls to reach PER=ρ per power gap applied to the nearest half of UEs to the BS in figure 3.4

independently of the channel conditions. This shows the relevance of power multiplexing in the uplink. This result also displays that equal cardinality between the high power set and low power set is a big contributer for the success of NOMA SIC removal. A 50%-50% distribution of the topology might not be so easy to attain in a multiple BS scenario. Such equal distribution results in the following cumulative distribution function at the receiver, depicted in figure 3.18. The vertical line in the CDF shows the perfect power separation at the BS provided by the fair power distribution between UEs.

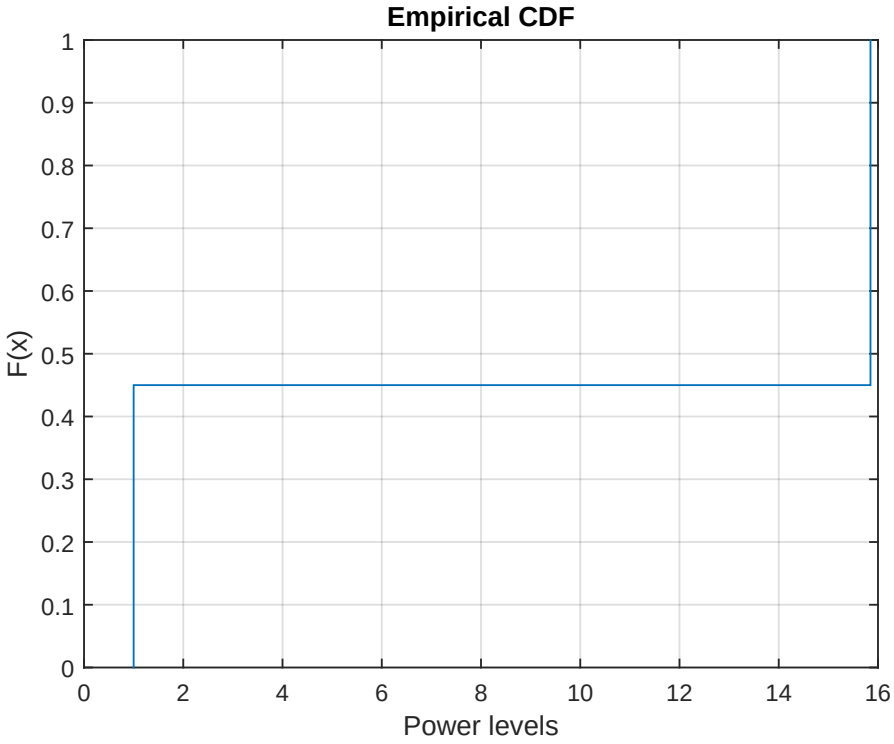


Figure 3.18: CDF (cumulative distribution function) of power levels at the receiver for a gap of 15dBs between power levels

Two BSs scenario

In a two BS framework, the high power subset H of UEs in a total set S of UEs served by its near BS, also impacts as interference in the farthest BS and, as a consequence, the 50%-50% distribution between high H UEs and low L UEs (with $\#H + \#L = \#S$) will no longer be achieved, and we therefore must stricly reinforce transmission power at the closest UEs and include absolute distance from UE p to BS j as a criterium in power scheduling.

Therefore, where and in which UEs must we employ high transmission power levels in topologies like those shown in figure 3.5 must be the subject of careful analysis when planning radio network performance tests. Although it is not the scope of this thesis to use algorihtms to distribute power levels throughout a network, optimal node diversity arrangement must be assumed if we want to study how to allocate workload between the nodes of a radio system.

In this subsection, two BSs arriving power perspectives are contrasted: 0 and 25dBs of power separation between UEs (i.e. without and with P-NOMA, respectively). The topology employed is that of figure 3.5 and optimal labelling between high and low UEs is needed to guarantee acceptable power separation and minimal interference. Thus, being BS_p the serving BS for UE_p, in a total set of J BSs in our topology, then UE_p can only belong to H_s, being H_s the set of the strongest UEs transmitting to its serving BS_p, if and only if (3.38) holds, for each j non-serving BS to UE_p in the system:

$$UE_p \in H_s \Leftrightarrow \Delta_{UE_p \longrightarrow BS_p} - \Delta_{UE_p \longrightarrow BS_j} \leq -3 \wedge \Delta_{UE_p \longrightarrow BS_p} < 3, \tag{3.38}$$

with $j \in J \wedge j \neq s$. Which means that the relative difference of distances of UE_p to BS_p and UE_p to other BS j in the system has to be smaller in value to -3m, and also the absolute distance for UE_p to BS_p must be less than 3m. Thus, UEs with the indexes 7, 16 and 26 constitute the high power subset H_s for BS 1 and 20, 6, 29, 27, 21 and 13 are the H_s subset chosen for BS 2. The disparity in power level distribution between both scenarios is best observed in figures 3.19 and 3.20. Each x value on the axis of the CDF plots represents the difference between the high and low power levels (converted to linear scale before the subtraction between high and low) as the high level is increased 1dB. In figure 3.19 it is possible to see that for the no-NOMA case, there is a 50%-50% distribution of UEs per BSs. But for the P-NOMA case, where criterion (3.38) was applied, there is a 30%-70% distribution of the high and low power levels for each subset of served UEs for each BS with a special higher rise in probability for the BS 1 curve due to bigger cardinality of the set of H_s for BS 2 interfering with BS 1.

The simulations worked as follows for the two scenarios: firstly each BS tries to solve all the 33 nodes topology. Then one by one, the weakest UE with respect to its signal power at said BS of the previous experiments is droped from resolution and its power is counted only as interference - following the model in section 3.5 - until each BS has only one UE to solve and 32 interfering UEs. The output is in the form of copies and computation time needed to solve a matrix with x UEs and 33-x interfering UEs. For

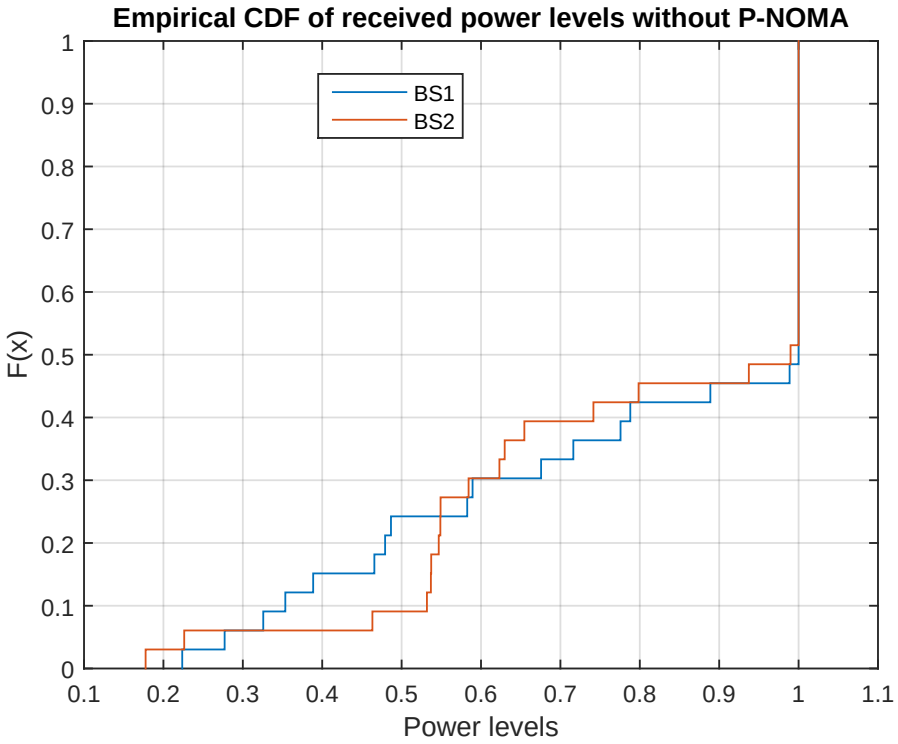


Figure 3.19: CDF (cumulative distribution function) for 2BSs scenario without P-NOMA.

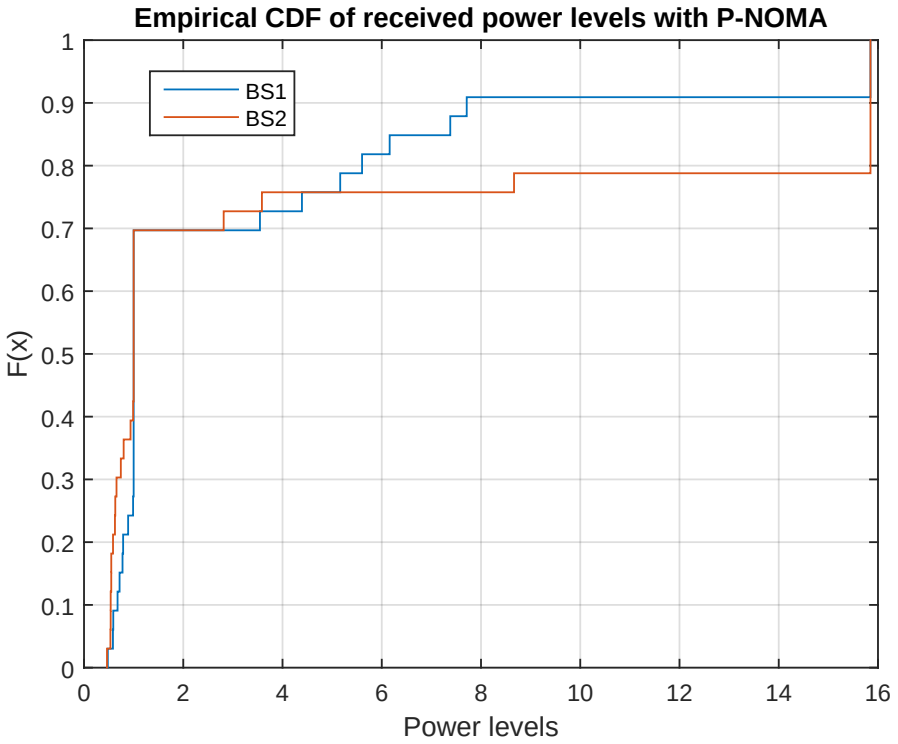


Figure 3.20: CDF (cumulative distribution function) for 2BSs scenario with P-NOMA.

time efficiency purposes, a simulation sealing of 64 copies of a packet is provided to each iteration of the test. Which means that for any given x UEs being solved, if the number of Ls (figures 3.21 and 3.23) to solve said x UEs equals 64, there is no resolution at that point x (PER of the UE with the highest PER value $\geq \rho$), thus x becomes $x_{sealing}$ and the simulation tests a bigger amount of UEs from left to right (or a smaller amount of UEs from right to left). At that point $x_{sealing}$, computation time values $y(x_{sealing})$ (figures 3.22 and 3.24) become useless for the study as well. The goal here is to investigate the moment where there might be a sharp performance improvement (a sudden drop in L copies or computation time needed) that might serve us as an evidence of the existence of an estimation metric that might function as a flag in order to best balance the workloads between the network.

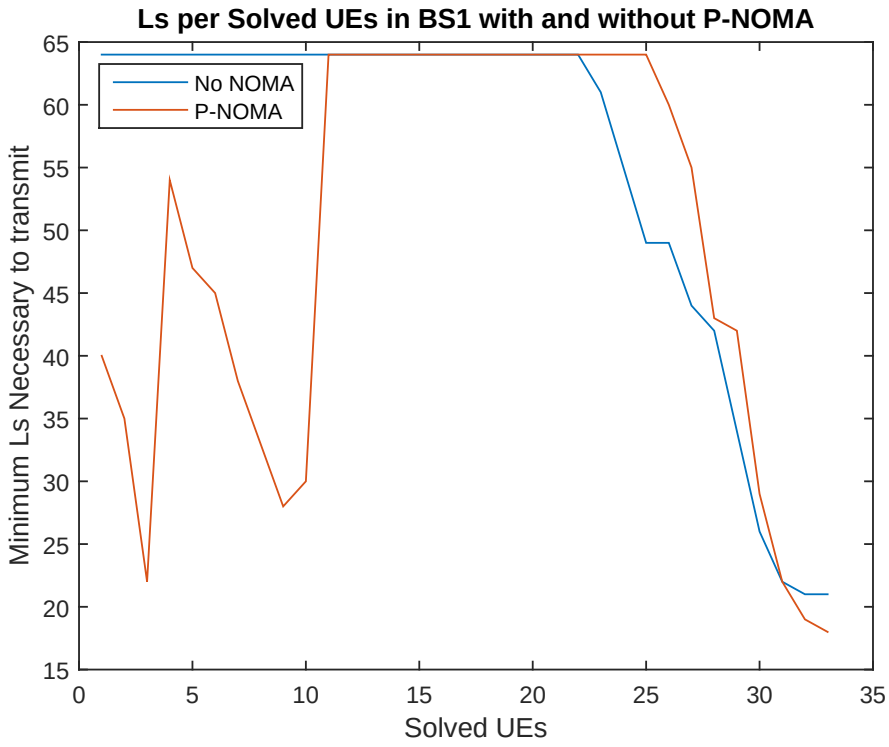


Figure 3.21: L (number of copies) needed to solve x UEs with 33-x interfering UEs for BS1 with and without P-NOMA.

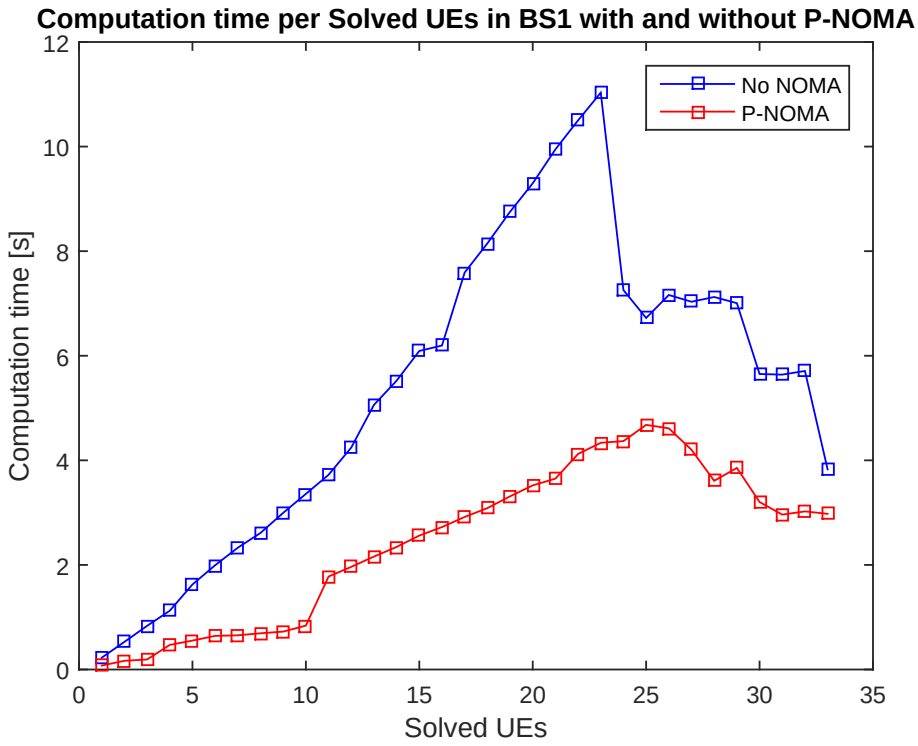


Figure 3.22: Computation time needed to solve x UEs with 33-x interfering UEs for BS1 with and without P-NOMA.

For these first two figures (3.21 and 3.22) pertaining to BS 1 in figure 3.5, we can already detect the better performance with respect to the P-NOMA strategy in comparison to the no-NOMA approach. The L curve reaches its sealing with better performance initially for the no-NOMA aproach (somewhere at the 21 UEs being solved for no-NOMA experiment and 25 UEs for the P-NOMA scheme), which means that initially (the term "initially"here means at the most right values of figure 3.21 and 3.22, from 33 to 25 solved UEs approximately), the no-NOMA scheme tolerates more interference for the same amount of copies needed for an overall system PER of ρ. It is observed that somewhere at the $8th$ simulation point where only 8 UEs are being solved, the performance increases significantly both for retransmissions needed in the matrix and also computation time at the receiver when power multiplexing is applied as the number of necessary copies of a packet to reach a PER of ρ sharpenly drops to 29. This might be a sign of sudden SINR improvement in reception due to only low power interfering UEs and only high power UEs being solved.

It is also verified in figure 3.22 a strange phenomenon since computation time needed to solve a matrix with approximately 23 users is higher than all the other matrices with larger number of users to solve, for both P-NOMA and no-NOMA approaches. Regarding this behavior, the author presumes that the interfence model adds a tradeoff margin to the system, since that, as users increase in the matrix, the number of interferents decrease in the experiment.

The subsequent two figures (3.23 and 3.24), reinforce the same results in the previous comparisons observed for BS 1 where at x=8 solved UEs there is an obvious increase in performance both in copies needed for each UE and computation time with respect to the P-NOMA approach. These tests open the possibility for an hierarchical solving scheme where x UEs can be solved locally at each BS and the joint processing CoMP matrix can be employed for inter-cell UEs.

3.6.3.4 SINR as an estimation measure for traffic balancing

From the tests made in section 3.6.3.3, it is noticeable that for a P-NOMA scheme, the gains at local (partial) processing can be attained as long as good network power scheduling between high power and lower power nodes is provided. On this subsection, SINR monitoring is made in order to better understand if, to each iteration of the simulations conducted in section 3.6.3.3, this performance metric can indicate relevant information about the possibility of slicing the workload of the system, thus decreasing delay in the network.

Since there is no analytical model to define SINR for the IB-DFE receptor used throughout the thesis, approximated calculations from previous works in the area were inspected in order to allow us to predict the behaviour of the system wihtout the need for exhaustive simulations. An estimation for SINR is proposed, based on the work in [35] that calculates SINR for a linear MMSE (Minimum Mean Square Error) receiver. According

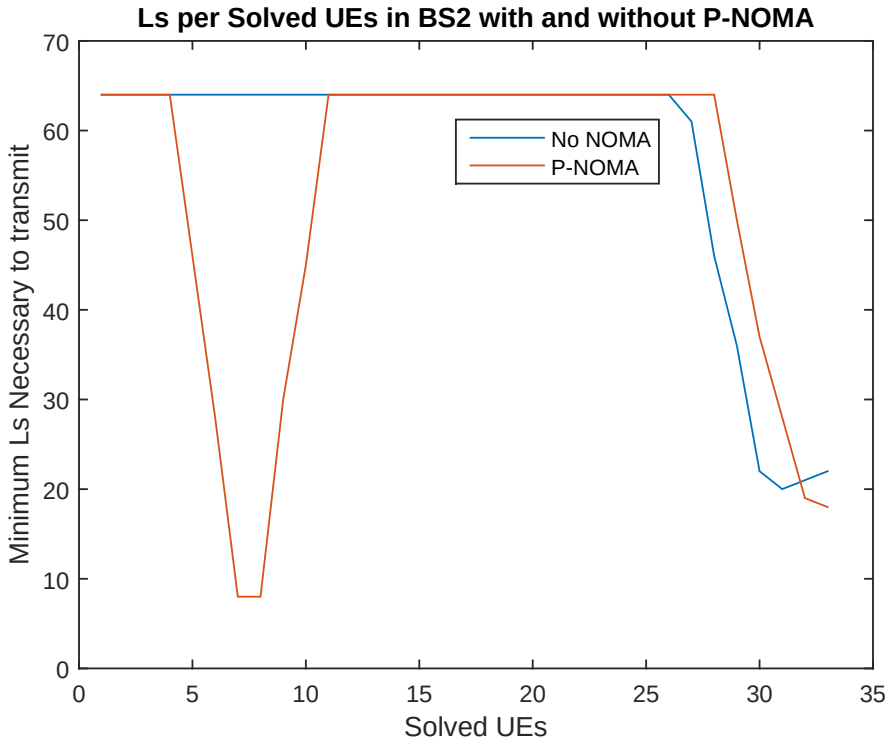


Figure 3.23: L (number of copies) needed to solve x UEs with 33-x interfering UEs for BS2 with and without P-NOMA.

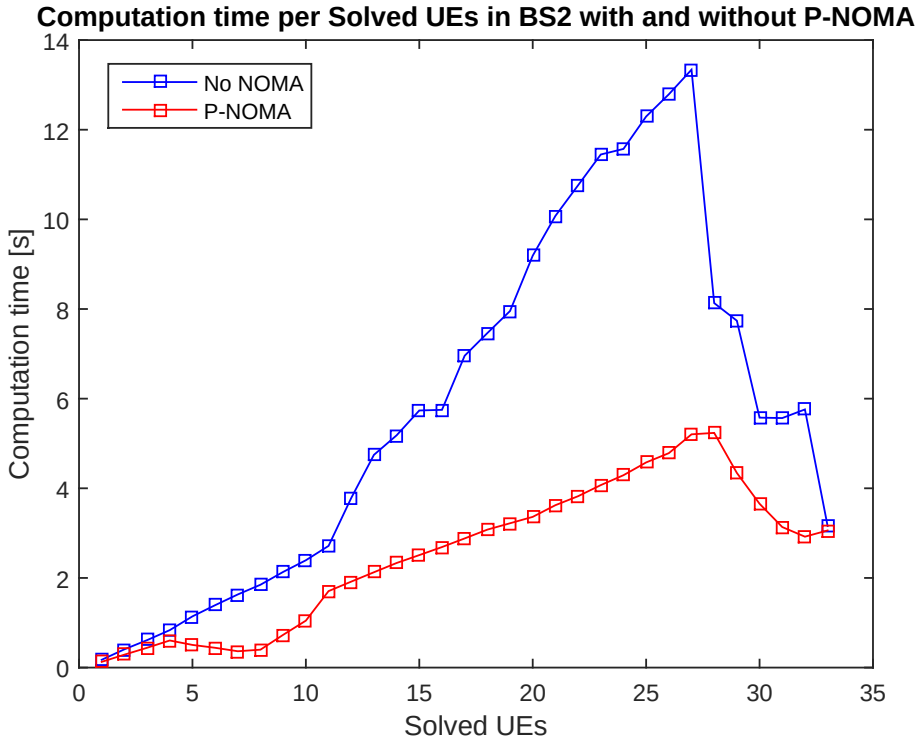


Figure 3.24: Computation time needed to solve x UEs with 33-x interfering UEs for BS2 with and without P-NOMA.

to theorem 3.1 in [35], let $\beta_1^{(L)}$ be the (random) SIR of a linear MMSE receiver for user 1 when the spreading length is (L). Then $\beta_1^{(L)}$ converges to β_1^* in probability as $L \to \infty$, where β_1^* is the unique solution to the equation

$$\beta_1^* = \frac{P_1}{\sigma^2 + \alpha \mathbb{E}_P\left[I(P, P_1, \beta_1^*)\right]}, \tag{3.39}$$

and

$$I(P, P_1, \beta_1^*) \equiv \frac{P P_1}{P_1 + P\beta_1^*}. \tag{3.40}$$

In (3.39), $\mathbb{E}_P[\cdot]$ denotes taking the expectation with respect to the limiting empirical distribution F of the received powers of the interferers. Heuristically, this means that in a large system, the SIR β_1 is deterministic and approximately satisfies

$$\beta_1 \approx \frac{P_1}{\sigma^2 + \frac{1}{L}\sum_{i=2}^{K} I(P_i, P_1, \beta_1)}, \tag{3.41}$$

where, P_i is the received power of user i. This result yields an interesting interpretation of the effect of each of the interfering users on the SINR of user 1: for a large system, the total interference can be decoupled into a sum of the background noise and an interference term from each of the other users (the factor $\frac{1}{L}$ results from the processing gain of user 1, and in this dissertation N equals number of copies L). Using the expression in (3.41) it is possible to obtain the following picture,

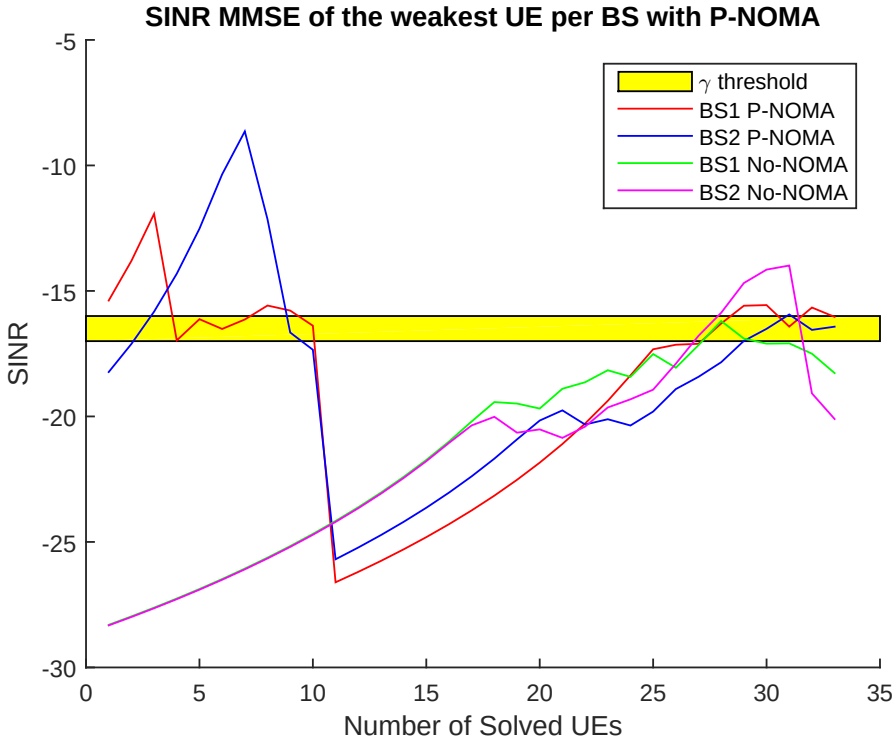


Figure 3.25: SINR of the weakest UEs for both BSs in figure 3.5 with and without P-NOMA using expression (3.41)

For each datapoint in figure 3.25, what is shown represents the monitoring of the SINR of the weakest UE being solved at each iteration of the tests made in 3.6.3.3. From right to left, as the amount of 33-x interfering UEs increases, the number of x solved UEs decreases, and the weakest UE from a given subset of S solved UEs of a given iteration n of a test is the next to count as interferent in the *n-1* iteration (if we subtract iterations from right to left). The value of this weakest UE is displayed in figure 3.25 and its variations in SINR are displayed.

As before, $SINR(x_{sealing})$ values for BS 1 and BS 2 can be dismissed, which are the same as in the tests for section 3.6.3.3 (for BS 1 $11 \leq x_{sealing} \leq 24$ and for BS 2 $11 \leq x_{sealing} \leq 26$, approximately).

Upon close inspection of non $x_{sealing}$ SINR values of the figure, we can see that there is a threshold γ of values between -16 and -17 dBs that defines the boundaries for the values of x (number of UEs received) where the L_s limit was reached (i.e. the $x_{sealing}$ limits), observed in figures 3.21 and 3.23. For SINR values above this threshold, the PER condition was verified with L_s values below 90. Reception is guaranteed if the value of SINR is superior to γ because at that given point x, $Max(PER(x)) \leq \rho$, using L_s copies for x UEs being solved. As in section 3.6.3.3 the P-NOMA experiment surpasses the no-NOMA approach at $x = 11$ approximately, where SINR values rise again to reach numbers near threshold γ.

What the figure 3.25 shows us, more that a gain in using P-NOMA, is the possibility to, given any configuration of UEs and BSs in a topology, and knowing the power scheduling and transmission powers of the nodes in our system, decide when and where processing allocation can be made in a more immediate fashion, allowing us to reduce latency in a network through the use of threshold γ as a criterion. Of course, the nature of the iterative receiver as a contribution for such a low γ value cannot be ignored. As shown in figure 3.3, the use of a generous amount of iterations (in this tests, the number of iterations used at the receiver is four), allows us to succeed in more restrictive energy consumption environments, which very well suit the context of MTC. Therefore MMSE SINR estimation can be a fairly reliable metric to subdivide processing demands throughout the network nodes in order to achieve our goal of URLLC requirements, due to the fact that it gives the author a legitimate criterion to use when deciding where to make matrices smaller in solving nodes, and decide to assign processing needs from local BSs to the cloud pool or vice-versa.

Chapter 4

Conclusions

4.1 Final Considerations

This dissertation studies the different uplink diversity techniques in a radio network, in an effort to efficiently combine various approaches to reduce overall latency of a given mobile communication system. Diversity schemes like Hybrid-CoMP and Power-NOMA are extensively explored in the dissertation to evaluate the possibility of distributing processing needs over a C-RAN network, which in turn, constitutes a new element of the fifth generations mobile networks and can be used as a tool to increase processing capability, mitigate latency and improve reliability in 5G mobile traffic. These diversity techniques all work on top of the implementation of IB-DFE receivers throughout the topology. These types of receivers employ an iterative process that allows the signals from various UEs to be solved under more restrictive energy requirements, which helps improving power saving, and as such, it is suitable for Machine Type Communication traffic.

The author, observing the output of simulations employing diversity approaches through metrics as computation time to solve a given matrix of N UEs, number of re-transmissions necessary to reach a certain PER threshold ρ and SINR at the receiver when applying power diversity, comes to the conclusion, by obtaining a SINR threshold γ where reception for higher power UEs is guaranteed, that under certain power configurations, in an 2 BS environment, the processing needs of the near BS UEs and cell-edge UEs can be separated efficiently in order to reduce latency in the network.

4.2 Future Work

In subsequent studies to follow the work presented on this dissertation, besides the minimum value γ obtained to guarantee reception at the receiver, studies that include more than 2 BSs and 2 power levels (which were the degrees of freedom considered on this thesis) can be examined to reiterate the validity of this SINR threshold.

Also, an algorithm can be built so that processing needs are allocated immediately (through C-RAN or local nodes), just by inspecting the SINR at each receiver, to decide where in the topology a given UE is going to be solved, given the minimum SINR value of a set of UEs. This type of quick decision-making in the network could be investigated in order to greatly reduce unecessary workload on each node of a mobile system.

Bibliography

[1] X. Li, J. B. Rao, and H. Zhang. "Engineering Machine-to-Machine Traffic in 5G." In: *IEEE Internet of Things Journal* 3.4 (2016), pp. 609–618.

[2] P. Kwadwo Agyapong, M. Iwamura, D. Staehle, W. Kiess, and A. Benjebbour. "Design Considerations for a 5G Network Architecture." In: *IEEE Communications Magazine* 52.11 (2014), pp. 1–16.

[3] J. Wu, Z. Zhang, Y. Hong, and Y. Wen. "Cloud Radio Access Network (C-RAN): A Primer." In: *IEEE Network* 29.1 (2015), pp. 35–41.

[4] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim. "Introduction to Ultra Reliable and Low Latency Communications in 5G." In: *IEEE Wireless Communications* 25.3 (2017).

[5] M. Richart, J. Baliosian, J. Serrat, and J. L. Gorricho. "Resource Slicing in Virtual Wireless Networks: A Survey." In: *IEEE Transactions on Network and Service Management* 13.3 (2016), pp. 462–476.

[6] M. Arslan, K. Sundaresan, and S. Rangarajan. "Software-defined networking in cellular radio access networks: Potential and challenges." In: *IEEE Communications Magazine* 53.1 (2015), pp. 150–156.

[7] X. Wang, S. Thota, M. Tornatore, H. S. Chung, H. H. Lee, S. Park, and B. Mukherjee. "Energy-Efficient Virtual Base Station Formation in Optical-Access-Enabled Cloud-RAN." In: *IEEE Journal on Selected Areas in Communications* 34.5 (2016), pp. 1130–1139.

[8] Shao-Yu Lien, Shao-Chou Hung, Kwang-Cheng Chen and Y.-C. Liang. "Ultra-Low-Latency Ubiquitous Connections in Heterogeneous Cloud Radio Access Networks." In: *IEEE Wirel Comm Mag* 22.3 (2015), pp. 22–31.

[9] D. Boviz, N. Abbas, G. Aravinthan, C. S. Chen, D. Boviz, N. Abbas, G. Aravinthan, C. S. Chen, and M. A. Dridi. "Multi-cell Coordination in Cloud RAN : Architecture and Optimization." In: *The international conference on wireless networks and mobile communications (WINCOM'16)* (2016).

[10] V. Naware and P. Venkitasubramaniam. "A cross-layer perspective in an uncharted path - Signal processing in random access." In: *IEEE Signal Processing Magazine* 21.5 (2004), pp. 29–39.

[11] M. Al-Imari, P. Xiao, M. A. Imran, and R. Tafazolli. "Uplink Non-Orthogonal Multiple Access for 5G Wireless Networks." In: *2014 11th International Symposium on Wireless Communications Systems (ISWCS)* (2014), pp. 781–785.

[12] R. Razavi, M. Dianati, and M. A. Imran. "Non-Orthogonal Multiple Access (NOMA) for future radio access." In: *5G Mobile Communications* (2016), pp. 135–163.

[13] L. Dai, B. Wang, Y. Yuan, S. Han, C. L. I, and Z. Wang. "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends." In: *IEEE Communications Magazine* 53.9 (2015), pp. 74–81.

[14] Z. Yang, Z. Ding, P. Fan, and N. Al-Dhahir. "A General Power Allocation Scheme to Guarantee Quality of Service in Downlink and Uplink NOMA Systems." In: *IEEE Transactions on Wireless Communications* 15.11 (2016), pp. 7244–7257.

[15] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura. "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)." In: *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC*. 2013, pp. 611–615.

[16] Z. Ding, P. Fan, and H. V. Poor. "Impact of User Pairing on 5G Nonorthogonal Multiple-Access Downlink Transmissions." In: *IEEE Transactions on Vehicular Technology* 65.8 (2016), pp. 6010–6023.

[17] Z. Ding, Z. Yang, P. Fan, and H. V. Poor. "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users." In: *IEEE Signal Processing Letters* 21.12 (2014), pp. 1501–1505.

[18] C. Xu, L. Ping, P. Wang, S. Chan, and X. Lin. "Decentralized power control for random access with successive interference cancellation." In: *IEEE Journal on Selected Areas in Communications* 31.11 (2013), pp. 2387–2396.

[19] Y. Liang, X. Li, J. Zhang, and Z. Ding. "Non-Orthogonal Random Access (NORA) for 5G Networks." In: *IEEE Transactions on Wireless Communications* 16.7 (2017), pp. 19–22.

[20] Y. Tian, A. R. Nix, and M. Beach. "On the Performance of Opportunistic NOMA in Downlink CoMP Networks." In: *IEEE Communications Letters* 20.5 (2016), pp. 998–1001.

[21] M. S. Ali, E. Hossain, and D. I. Kim. "Coordinated Multi-Point Transmission in Downlink Multi-cell NOMA Systems: Models and Spectral Efficiency Performance." In: *IEEE Wireless Communications* 25.5 (2017), pp. 24–31.

[22] M. S. Ali, H. Tabassum, and E. Hossain. "Dynamic User Clustering and Power Allocation for Uplink and Downlink Non-Orthogonal Multiple Access (NOMA) Systems." In: *IEEE Access* 4 (2016), pp. 6325–6343.

[23] Q.-t. Vien, N. Ogbonna, H. X. Nguyen, R. Trestian, and P. Shah. "On the Non-Orthogonal Multiple Access for Downlink in Cloud Radio Access Networks On the Non-Orthogonal Multiple Access for Downlink in Cloud Radio Access Networks." In: *Technical Report* February 2016 (2014). URL: https://www.researchgate.net/publication/266742238_On_the_Non-Orthogonal_Multiple_Access_for_Downlink_in_Cloud_Radio_Access_Networks.

[24] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. Bhargava. "A Survey on Non-Orthogonal Multiple Access for 5G Networks: Research Challenges and Future Trends." In: *IEEE Journal on Selected Areas in Communications* 35.10 (2017), pp. 1–32.

[25] Y. Du, B. Dong, Z. Chen, J. Fang, and L. Yang. "Shuffled Multiuser Detection Schemes for Uplink Sparse Code Multiple Access Systems." In: *IEEE Communications Letters* 20.6 (2016), pp. 1231–1234.

[26] F. Ganhão, M. Pereira, L. Bernardo, R. Dinis, R. Oliveira, and P. Pinto. "Performance analysis of an hybrid ARQ adaptation of NDMA schemes." In: *IEEE Transactions on Communications* 61.8 (2013), pp. 3304–3317.

[27] M. K. Tsatsanis, R. Zhang, and S. Banerjee. "Network-assisted diversity for random access wireless networks." In: *IEEE Transactions on Signal Processing* 48.3 (2000), pp. 702–711.

[28] F. Ganhão, R. Dinis, L. Bernardo, and R. Oliveira. "Analytical BER and PER performance of frequency-domain diversity combining, multipacket detection and hybrid schemes." In: *IEEE Transactions on Communications* 60.8 (2012), pp. 2353–2362.

[29] B. Ramos, L. Bernardo, R. Dinis, R. Oliveira, P. Pinto, and P. Amaral. "Using lightly synchronized MultiPacket Reception in Machine-Type Communication networks." In: *2016 IEEE Globecom Workshops, GC Wkshps 2016 - Proceedings* (2016).

[30] R. Dinis, P. Silva, and A. Gusmão. "IB-DFE receivers with space diversity for CP-assisted DS-CDMA and MC-CDMA systems." In: *European Transactions on Telecommunications* 18.7 (2007), pp. 791–802.

[31] A. Gusmão, P. Torres, R. Dinis, and N. Esteves. "A turbo FDE technique for reduced-CP SC-based block transmission systems." In: *IEEE Transactions on Communications* 55.1 (2007), pp. 16–20.

[32] A. B. Carlson and P. B. Crilly. *Communication Systems, 4th Ed.* 2001, McGraw Hill Higher Education. ISBN: 978–0071210287.

[33] R. Dinis, P. Montezuma, N. Souto, and J. Silva. "Iterative frequency-domain equalization for general constellations." In: *33rd IEEE Sarnoff Symposium 2010, Conference Proceedings*. 2010.

[34] R. Dinis, P. Montezuma, L. Bernardo, R. Oliveira, M. Pereira, and P. Pinto. "Frequency-domain multipacket detection: A high throughput technique for SC-FDE sstems." In: *IEEE Transactions on Wireless Communications* 8.7 (2009), pp. 3798–3807.

[35] D Tse and S Hanly. "Multi-user Demodulation: Effective Interference, Effective Bandwidth and Capacity." In: *IEEE Transactions on Information Theory* 45.2 (1999), pp. 641–657.